A Stochastic Path-Integrated Differential EstimatoR Expectation Maximization Algorithm

Gersende Fort, Eric Moulines, Hoi-To Wai

A Stochastic Path-Integrated Differential EstimatoR Expectation Maximization Algorithm

Gersende Fort
Institut de Mathématiques de Toulouse
Université de Toulouse; CNRS
UPS, Toulouse, France
gersende.fort@math.univ-toulouse.fr

Eric Moulines
Centre de Mathématiques Appliquées
Ecole Polytechnique, France
CS Dpt, HSE University, Russian Federation
eric.moulines@polytechnique.edu

Hoi-To Wai
Department of SEEM
The Chinese University of Hong Kong
Shatin, Hong Kong
htwai@cuhk.edu.hk

Abstract

The Expectation Maximization (EM) algorithm is of key importance for inference in latent variable models including mixture of regressors and experts, missing observations. This paper introduces a novel EM algorithm, called SPIDER-EM, for inference from a training set of size n, $n \gg 1$. At the core of our algorithm is an estimator of the full conditional expectation in the E-step, adapted from the stochastic path-integrated differential estimator (SPIDER) technique. We derive finite-time complexity bounds for smooth non-convex likelihood: we show that for convergence to an ϵ-approximate stationary point, the complexity scales as $K_{\mathrm{Opt}}(n,\epsilon) = \mathcal{O}(\epsilon^{-1})$ and $K_{\mathrm{CE}}(n,\epsilon) = n + \sqrt{n}\mathcal{O}(\epsilon^{-1})$, where $K_{\mathrm{Opt}}(n,\epsilon)$ and $K_{\mathrm{CE}}(n,\epsilon)$ are respectively the number of M-steps and the number of per-sample conditional expectations evaluations. This improves over the state-of-the-art algorithms. Numerical results support our findings.

This paper is close to the final version accepted for publication in the Conference on Neural Information Processing Systems (NeurIPS 2020). The final version can be found at
https://papers.nips.cc/paper/2020/hash/c589c3a8f99401b24b9380e86d939842-Abstract.html

1 Introduction

Expectation Maximization (EM) is a key algorithm in machine-learning and statistics [20]. Applications are numerous including clustering, natural language processing, parameter estimation in mixed models, missing data, to give just a few. The common feature of all these applications is the introduction of latent variables: the "incomplete" likelihood $p(y;\theta)$ where $\theta \in \Theta \subseteq \mathbb{R}^d$ is defined by marginalizing the "complete-data" likelihood $p(y,z;\theta)$ defined as the joint distribution of the observation y and a non-observed latent variable $z \in \mathsf{Z}$, i.e. $p(y;\theta) = \int p(y,z;\theta)\mu(\mathrm{d}z)$ where Z is the latent space and μ is a measure on Z. We focus in this paper on the case where $p(y,z;\theta)$ belongs to a curved exponential family, given by

$$p(y,z;\theta) \stackrel{\mathrm{def}}{=} \rho(y,z)\exp\left\{\langle s(y,z),\phi(\theta)\rangle - \psi(\theta)\right\}; \tag{1}$$

where $s(y,z) \in \mathbb{R}^q$ is the complete data sufficient statistics, $\phi : \Theta \to \mathbb{R}^q$ and $\psi : \Theta \to \mathbb{R}$, $\rho : \mathsf{Y} \times \mathsf{Z} \to \mathbb{R}^+$ are vector/scalar functions. Given a training set of n independent observations

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

$\{y_i\}_{i=1}^n$, our goal is to minimize the negated penalized log-likelihood with respect to $\theta \in \Theta$:

$$\min_{\theta\in\Theta} F(\theta) \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}_i(\theta) + \mathsf{R}(\theta), \quad \mathcal{L}_i(\theta) \stackrel{\text{def}}{=} -\log p(y_i;\theta), \tag{2}$$

such that $\mathsf{R}(\theta)$ is a regularizer. A popular solution approach to (2) is the EM algorithm [10] which is a special instance of the Majorize-Minimization (MM) algorithm. It alternates between two steps: in the Expectation (E) step, using the current value of the iterate θ_{curr}, we compute a majorizing function $\theta \mapsto \mathsf{Q}(\theta, \theta_{\text{curr}})$ given up to an additive constant by

$$\mathsf{Q}(\theta, \theta_{\text{curr}}) \stackrel{\text{def}}{=} -\langle \bar{s}(\theta_{\text{curr}}), \phi(\theta)\rangle + \psi(\theta) + \mathsf{R}(\theta) \quad \text{where} \quad \bar{s}(\theta) \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} \bar{s}_i(\theta)\ ; \tag{3}$$

and $\bar{s}_i(\theta)$ is the ith sample conditional expectation of the complete data sufficient statistics:

$$\bar{s}_i(\theta) \stackrel{\text{def}}{=} \int_{\mathsf{Z}} s(y_i, z) p(z|y_i;\theta)\mu(\mathrm{d}z)\ , \quad p(z|y_i;\theta) \stackrel{\text{def}}{=} p(y_i, z;\theta)/p(y_i;\theta)\ . \tag{4}$$

As for the Maximization (M) step, a new value of θ_{curr} is computed as a minimizer of $\theta \mapsto \mathsf{Q}(\theta, \theta_{\text{curr}})$. The majorizing function is then updated with the new θ_{curr}. This process is iterated until convergence. One of the distinctive advantage of EM algorithms with respect to (w.r.t.) first-order methods stems from the fact that it is invariant by change of parameterization and that EM is, by construction, monotone; see [20].

The conventional EM algorithm is not suitable for analyzing the increasingly large data sets, such as those that could be considered as big data in volumes [5, 14]: in such case, the explicit computation of $\bar{s}(\theta_{\text{curr}})$ in *each* E*-step* of the EM algorithm involves evaluating n conditional expectations [20]. As a remedy, *incremental* methods were designed which reduce the number of samples used per iteration to a mini-batch. Among the incremental methods, the first approach to cope with large-scale EM setting is the incremental EM (`iEM`) algorithm [21] (also see [22] for a refined algorithm). At each iteration, `iEM` selects a minibatch $\mathcal{B}_{\text{curr}}$ of size b and updates the associated statistic $\bar{s}_i(\theta_{\text{curr}}), i \in \mathcal{B}_{\text{curr}}$, in the current estimate $\widehat{S}_{\text{curr}}$ of $\bar{s}(\theta_{\text{curr}})$; and then updates the parameters by a classical M-step. Later, an alternative approach was proposed in [6] as the `Online EM` algorithm, which shares some similarities with stochastic gradient descent [4] even though `Online EM` is not a first-order method. Recent papers have proposed improvements to `Online EM` by combining it with variance reduction techniques. For instance, [7] and [18] proposed respectively the stochastic EM with variance reduction (`sEM-vr`) and the fast incremental EM (`FIEM`) algorithms. These methods are extensions to the EM algorithm of the SVRG [15] and the SAGA [8] techniques.

The complexity of these algorithms have been analyzed under the assumption that $F(\theta)$ is smooth but possibly non-convex. They are expressed as the number of M-steps updates, $K_{\text{Opt}}(n, \epsilon)$, and the number of per-sample conditional expectations evaluations $K_{\text{CE}}(n, \epsilon)$, in order to find an ϵ-approximate stationary point of $F(\theta)$; see (11) for the definition. It was established in [18] that $K_{\text{Opt}}(n, \epsilon) = K_{\text{CE}}(n, \epsilon) = n + n^{2/3}\mathcal{O}(\epsilon^{-1})$ updates/evaluations are needed for the `sEM-vr` and `FIEM` algorithms (the rate for `FIEM` can be sharpened, see [12]). These complexity bounds match those of the SVRG and the SAGA algorithms for smooth non-convex optimization [25].

For smooth non-convex problems, the Stochastic Path-Integrated Differential EstimatoR (`SPIDER`) technique has recently been introduced by [11] (see also [27] for `SPIDER-BOOST` and [24] for `SARAH`), which established an $n + \sqrt{n}\mathcal{O}(\epsilon^{-1})$ bound of calls to first order oracles to find an ϵ-approximate stationary solution of a general finite sum optimization problem. Furthermore, the $\sqrt{n}$-dependence was proven to be optimal. This motivates the current work to explore new EM algorithms with reduced complexity. Our contributions are:

- We propose a novel `SPIDER-EM` algorithm, inspired by the `SPIDER` estimator in [11] and tailored to the EM framework for curved exponential family class of distributions. The `SPIDER-EM` uses an outer loop to maintain a *control variate* that requires a full scan of the dataset to compute $\bar{s}(\theta_{\text{curr}})$, and inner loops which perform low complexity updates by drawing random minibatches of samples.
- We introduce a unified framework of *stochastic approximation (SA) within EM* which covers the convergence analysis of `Online EM`, `sEM-vr`, `FIEM`, `SPIDER-EM`. In this general framework, `SPIDER-EM` may be seen as a stochastic approximation algorithm using variance reduced estimate $\widehat{S}_{\text{curr}}$.

- Using the SA analysis framework, we prove that the complexity bounds for `SPIDER-EM` are $K_{\mathrm{Opt}}(n,\epsilon) = \mathcal{O}(\epsilon^{-1})$, $K_{\mathrm{CE}}(n,\epsilon) = n + \sqrt{n}\mathcal{O}(\epsilon^{-1})$. Among the incremental-EM techniques, we provide state of the art complexity bounds that overpass all the previous ones.
- The EM is not a first-order method contrary to `SPIDER`. Therefore, the convergence analysis of `SPIDER-EM` methods require specific mathematical developments which differ significantly from the original `SPIDER` analysis. In addition, the analysis of `SPIDER-EM` differs from previous ones for incremental EM algorithms, since it involves *biased* approximations, which makes the proof more challenging (see section 9, Lemma 11).
- We provide a new perspective to interpret `SPIDER-EM` as an equivalent algorithm to a perturbed `Online-EM` where the perturbation acts as a control variate to reduce variance - see algorithm 7.

Furthermore, the `SPIDER-EM` algorithm operates with a significantly lower memory footprint than `iEM` and `FIEM`, and the memory footprint is on par with `sEM-vr` and `Online EM`. To our best knowledge, the proposed algorithm offers the best of both worlds – having a low complexity bounds and a low memory footprint. Lastly, we support the theoretical findings with numerical experiments and show that `SPIDER-EM` performs favorably compared to existing algorithms.

Notations. For two vectors $a, b \in \mathbb{R}^r$, $\langle a, b\rangle$ denotes the usual Euclidean product and $\|a\|$ the associated norm. By convention, vectors are column vectors. For a vector x with components $(x_1, \ldots, x_r)$, $x_{i:j}$ denotes the sub-vector with components $(x_i, x_{i+1}, \ldots, x_{j-1}, x_j)$. For two matrices $A \in \mathbb{R}^{r_1 \times r_2}$ and $B \in \mathbb{R}^{r_3 \times r_4}$, $A \otimes B$ denotes the Kronecker product. I_r is the $r \times r$ identity matrix. A^T is the transpose of A.

2 EM Algorithm and its Variants using Stochastic Approximation

We formulate the model assumptions and introduce the `SPIDER-EM` algorithm. Recall the definition of the negated penalized log-likelihood $F(\theta)$ from (2) and consider a few regulatory assumptions:

H1. *$\Theta \subseteq \mathbb{R}^d$ is a measurable convex set. $(\mathsf{Z}, \mathcal{Z})$ is a measurable space and μ is a σ-finite positive measure on $\mathcal{Z}$. The functions $\mathsf{R} : \Theta \to \mathbb{R}$, $\phi : \Theta \to \mathbb{R}^q$, $\psi : \Theta \to \mathbb{R}$, and $\rho(y_i, \cdot) : \mathsf{Z} \to \mathbb{R}_+$, $s(y_i, \cdot) : \mathsf{Z} \to \mathbb{R}^q$ for $i \in \{1, \ldots, n\}$ are measurable functions. For any $\theta \in \Theta$ and $i \in \{1, \ldots, n\}$, the log-likelihood is bounded as $-\infty < \mathcal{L}_i(\theta) < \infty$.*

H2. *For all $\theta \in \Theta$ and $i \in \{1, \ldots, n\}$, the conditional expectation $\bar{s}_i(\theta)$ is well-defined.*

H3. *For any $s \in \mathbb{R}^q$, the map $s \mapsto \mathrm{Argmin}_{\theta \in \Theta} \{\psi(\theta) + \mathsf{R}(\theta) - \langle s, \phi(\theta)\rangle\}$ exists and is unique; the singleton is denoted by $\{\mathsf{T}(s)\}$.*

As discussed in the Introduction, the EM algorithm is an MM algorithm associated with the majorization functions $\{\theta \mapsto \mathsf{Q}(\theta, \theta_{\mathrm{curr}}), \theta_{\mathrm{curr}} \in \Theta\}$. Thus, the EM algorithm defines a sequence $\{\theta_k, k \geq 0\}$ that can be computed recursively as $\theta_{k+1} = \mathsf{T} \circ \bar{s}(\theta_k)$, where the map T is defined in H3 and $\bar{s}$ is defined in (3). On the other hand, the EM algorithm can be defined through a mapping in the complete data sufficient statistics, referred to as the *expectation space*. In this setting, the EM iteration defines a sequence in $\mathbb{R}^q$ $\{\widehat{S}_k, k \geq 0\}$ given by $\widehat{S}_{k+1} = \bar{s} \circ \mathsf{T}(\widehat{S}_k)$. To summarize, we observe that the EM algorithm admits two equivalent representations:

$$\text{(Parameter space)}\ \ \theta_{k+1} = \mathsf{T} \circ \bar{s}(\theta_k); \quad \text{(Expectation space)}\ \ \widehat{S}_{k+1} = \bar{s} \circ \mathsf{T}(\widehat{S}_k). \tag{5}$$

In this paper, we focus on the expectation space representation. Let $\theta_\star \overset{\mathrm{def}}{=} \mathsf{T}(s_\star)$ where $s_\star \in \mathbb{R}^q$. It has been shown in [9] that if $s_\star$ is a fixed point to the EM algorithm in the expectation space, then $\theta_\star = \mathsf{T}(s_\star)$ is a fixed point of the EM algorithm in the parameter space, i.e., $\theta_\star = \mathsf{T} \circ \bar{s}(\theta_\star)$. Note that the converse is also true. The limit points of the EM algorithm in the expectation space are the roots of the *mean field*

$$h(s) \overset{\mathrm{def}}{=} \bar{s} \circ \mathsf{T}(s) - s, \quad s \in \mathbb{R}^q\,. \tag{6}$$

Consider the following assumption.

H4. *1. The functions ϕ, ψ and R are continuously differentiable on Θ^v. If Θ is open, then $\Theta^v = \Theta$, otherwise Θ^v is a neighborhood of Θ. T is continuously differentiable on $\mathbb{R}^q$.*

2. The function F is continuously differentiable on Θ^v and for any $\theta \in \Theta$, $\nabla F(\theta) = -\nabla\phi(\theta)^\top \bar{s}(\theta) + \nabla\psi(\theta) + \nabla\mathsf{R}(\theta)$.

3. For any $s \in \mathbb{R}^q$, $B(s) \stackrel{\text{def}}{=} \nabla(\phi \circ \mathsf{T})(s)$ is a symmetric matrix with positive minimal eigenvalue.

These assumptions are classical, see for example, [18] and the references therein.

A key property of the EM algorithm is that it is *monotone*: in the parameter space $\theta_{k+1} = \mathsf{T} \circ \bar{s}(\theta_k)$ decreases the objective function with $F(\theta_{k+1}) \leq F(\theta_k)$. The same monotone property also holds in the expectation space. Define

$$\mathrm{W}(s) \stackrel{\text{def}}{=} F \circ \mathsf{T}(s) = \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}_i(\mathsf{T}(s)) + \mathsf{R}(\mathsf{T}(s)), \quad s \in \mathbb{R}^q \,. \tag{7}$$

It can be shown that $F(\theta_{k+1}) \leq F(\theta_k)$ implies $\mathrm{W}(\hat{S}_{k+1}) \leq \mathrm{W}(\hat{S}_k)$. In addition, [9] showed that:

Proposition 1. *Under H1, H2, H3 and H4, $\mathrm{W}(s)$ is continuously differentiable on $\mathbb{R}^q$ and for any $s \in \mathbb{R}^q$, $\nabla \mathrm{W}(s) = -B(s)\, h(s)$.*

Hence, $s_\star$ is a fixed point to the EM algorithm in expectation space, with $s_\star = \bar{s} \circ \mathsf{T}(s_\star)$ and $h(s_\star) = 0$ if and only if $s_\star$ is a stationary point satisfying $\nabla \mathrm{W}(s_\star) = 0$. This property has made it possible to develop a new class of algorithms that preserve desirable properties of the EM (e.g, invariant in the choice of parameterization) while replacing the computation of $\bar{s}(\theta)$ by a stochastic approximation (SA) scheme; see [26, 2, 3] for a survey on SA. This scheme has been exploited in [9] to deal with the case where the computation of the conditional expectation $\bar{s}(\theta)$ is intractable.

We consider yet another form of intractability in this work which is linked with the size of the dataset $n \gg 1$. To alleviate this problem, the `Online EM` algorithm [6] defines a sequence $\{\widehat{S}_k, k \geq 0\}$ with the recursion:

$$\widehat{S}_{k+1} = \widehat{S}_k + \gamma_{k+1} \left(\bar{s}_{\mathcal{B}_{k+1}} \circ \mathsf{T}(\widehat{S}_k) - \widehat{S}_k \right) \,, \tag{8}$$

where $\{\gamma_{k+1}, k \geq 0\}$ is a deterministic sequence of step sizes, $\mathcal{B}_{k+1}$ is a mini-batch of b examples sampled at random in $\{1, \ldots, n\}$ and for a mini-batch $\mathcal{B}$ of size b, we set $\bar{s}_{\mathcal{B}} \stackrel{\text{def}}{=} \mathsf{b}^{-1} \sum_{i \in \mathcal{B}} \bar{s}_i$.

The `Online EM` algorithm can be viewed as an SA scheme designed for finding the roots of the mean-field h; indeed, the mean-field of `Online EM` satisfies $\mathbb{E}[\bar{s}_{\mathcal{B}_{k+1}} \circ \mathsf{T}(\widehat{S}_k) - \widehat{S}_k] = h(\widehat{S}_k)$. Hence, the possible limiting points of `Online EM` are the roots of $h(s)$, such a root $s_\star$ is a stationary point of W (see Proposition 1 and (7)), and $\mathsf{T}(s_\star)$ corresponds to a stationary point of the penalized likelihood (2); see [6] for a precise statement and [17] for a detailed convergence analysis.

Variance Reduction for SA with EM Algorithm. For the finite-sum problem (2), more efficient algorithms can be developed by introducing a control variate in order to achieve variance reduction. Suppose that we have a random variable (r.v.) U and our aim is to estimate $u \stackrel{\text{def}}{=} \mathbb{E}[U]$. For any zero-mean r.v. V, the sum $U + V$ is an unbiased estimator of u. Now, if V is negatively correlated with U and $\mathrm{Var}(V^2) \leq -2\,\mathrm{Cov}(U, V)$, then the variance of $U + V$ will be lower than that of the standalone estimator U; V is a *control variate*.

This approach has been proven to be effective for stochastic gradient algorithms: emblematic examples are Stochastic Variance Reduced Gradient (SVRG) introduced by [15] and SAGA introduced by [8]. Whereas control variates have been originally designed to the stochastic gradient framework, similar ideas can be applied to SA procedures for finite-sum optimization. For `Online EM`, variance reduction amounts to expressing the mean-field as $h(s) = \mathbb{E}\left[\bar{s}_{\mathcal{B}} \circ \mathsf{T}(s) - s + V\right]$ where V is a control variate. These methods differ in the way the control variate is constructed. The efficiency of such variance reduction methods improves with the correlation of V with $\bar{s}_{\mathcal{B}} \circ \mathsf{T}(s) - s$.

An SVRG-like algorithm is the `Stochastic EM with Variance Reduction (sEM-vr)` algorithm [7]. In `sEM-vr`, the control variate is reset in an outer loop every k_{in} iterations: in the outer loop $\#t$ for $t \in \{1, \ldots, k_{\text{out}}\}$, and the inner loop $\#(k+1)$ for $k \in \{0, \ldots, k_{\text{in}} - 2\}$, the complete data sufficient statistic is updated using `Online EM` and a recursively defined control variate

$$\widehat{S}_{t,k+1} = \widehat{S}_{t,k} + \gamma_{t,k+1}(\bar{s}_{\mathcal{B}_{t,k+1}} \circ \mathsf{T}(\widehat{S}_{t,k}) - \widehat{S}_{t,k} + V_{t,k+1}) \,, \tag{9}$$

$$V_{t,k+1} = \bar{s} \circ \mathsf{T}(\widehat{S}_{t-1,k_{\text{in}}-1}) - \bar{s}_{\mathcal{B}_{t,k+1}} \circ \mathsf{T}(\widehat{S}_{t-1,k_{\text{in}}-1}) \,. \tag{10}$$

When $k = 0$, the complete data sufficient statistic $\widehat{S}_{t,0}$ is obtained by performing first a full-pass on the dataset $\widetilde{S}_{t,0} = \bar{s} \circ \mathsf{T}(\widehat{S}_{t-1,k_{\text{in}}-1})$ and then updating $\widehat{S}_{t,0} = \widehat{S}_{t-1,k_{\text{in}}-1} + \gamma_{t,0}(\widetilde{S}_{t,0} - \widehat{S}_{t-1,k_{\text{in}}-1})$.

An SAGA-like version is the `Fast Incremental EM (FIEM)` algorithm proposed in [18]. The construction of the control variate for `FIEM` is more involved; for details, see algorithm 5 in the supplementary material.

In [18], the `sEM-VR` and `FIEM` algorithms have been analyzed with a randomized terminating iteration (τ, ξ), uniformly selected from $\{1, \ldots, k_{\text{out}}\} \times \{0, \ldots, k_{\text{in}} - 1\}$ where k_{in} (resp. k_{out}) is the number of inner loops per outer one, and k_{out} is the total number of outer loops. The random termination is inspired by [13] which enables one to show non-asymptotic convergence of stochastic gradient methods to a stationary point. Consider first `sEM-VR`. For any n, ϵ, we define $\mathcal{K}(n, \epsilon) \subset \mathbb{N}^3$ such that, for any $(k_{\text{in}}, k_{\text{out}}, \mathsf{b}) \in \mathcal{K}(n, \epsilon)$,

$$\mathbb{E}[\|h(\widehat{S}_{\tau,\xi})\|^2] \stackrel{\text{def}}{=} {k_{\max}}^{-1} \sum_{t=1}^{k_{\text{out}}} \sum_{k=0}^{k_{\text{in}}-1} \mathbb{E}[\|h(\widehat{S}_{t,k})\|^2] \leq \epsilon \,, \tag{11}$$

where $k_{\max} = k_{\text{in}} k_{\text{out}}$. In words, the randomly terminated algorithm computes a solution $\widehat{S}_{\tau,\xi}$ such that the expected squared norm of the mean field is less than ϵ; see [13]. The finite sample complexity in terms of the number of M-steps is $K_{\text{Opt}}^{\texttt{sEM-VR}}(n, \epsilon) = \inf_{\mathcal{K}(n,\epsilon)} k_{\text{in}} k_{\text{out}}$.

The complexity in terms of the total number of per-sample conditional expectations evaluations, is defined as $K_{\text{CE}}^{\texttt{sEM-VR}}(n, \epsilon, \mathsf{b}) = \inf_{\mathcal{K}(n,\epsilon)}\{n + k_{\text{out}} n + \mathsf{b} k_{\text{in}} k_{\text{out}} + (n \wedge (\mathsf{b} k_{\text{in}})) k_{\text{out}}\}$. Similar results can be derived for `FIEM` and other incremental EM algorithms (see section 6). In such case, define by $k_{\max} = k_{\max}(n, \epsilon)$ the minimal number of iterations such that (11) is satisfied and set $K_{\text{Opt}}^{\texttt{FIEM}}(n, \epsilon) = k_{\max}(n, \epsilon)$ and $K_{\text{CE}}^{\texttt{FIEM}}(n, \epsilon) = 2 k_{\max}(n, \epsilon) \mathsf{b}$. It can be shown (see [18] and the supplementary material) that $K_{\text{Opt}}^{\texttt{sEM-VR}}(n, \epsilon) = K_{\text{Opt}}^{\texttt{FIEM}}(n, \epsilon) = n^{2/3} \mathcal{O}(\epsilon^{-1})$ and $K_{\text{CE}}^{\texttt{sEM-VR}}(n, \epsilon) = K_{\text{CE}}^{\texttt{FIEM}}(n, \epsilon) = n + n^{2/3} \mathcal{O}(\epsilon^{-1})$. These bounds exhibit an $\mathcal{O}(\epsilon^{-1})$ growth as the stationarity requirement ϵ decreases. Such a rate is comparable to a deterministic gradient method for smooth and non-convex objective functions. However, the complexity of M-step computations as well as of conditional expectations evaluations grow at the rate of $n^{2/3}$, which can be undesirable if $n \gg 1$. Hereafter, we aim to design a novel algorithm with better finite-time complexities.

3 The `SPIDER-EM` Algorithm

To reduce the dependence on n and the overall complexity, we propose to design a *new control variate*, and to optimize the size of the *minibatch*. To this regard, we borrow from [11, 27] (see also [24] and the algorithm `SARAH`) a new technique called Stochastic Path-Integrated Differential Estimator (`SPIDER`) to generate the control variates for estimating the conditional expectation of the complete data for the full dataset.

Algorithm Description. We propose the `SPIDER-EM` algorithm formulated in the expectation space. The outer loop is the same as that of `sEM-vr`. The difference lays in the update of $\widehat{S}_k$ as follows:

Data: $k_{\text{in}} \in \mathbb{N}_\star$, $k_{\text{out}} \in \mathbb{N}_\star$, $\widehat{S}_{\text{init}} \in \mathbb{R}^q$, $\{\gamma_{t,k+1}, t \geq 1, k \geq 0\}$ positive sequence.
Result: The `SPIDER-EM` sequence: $\widehat{S}_{t,k}, t = 1, \ldots, k_{\text{out}}$ and $k = 0, \ldots, k_{\text{in}} - 1$

$\widehat{S}_{1,0} = \widehat{S}_{1,-1} = \widehat{S}_{\text{init}}, \quad \mathsf{S}_{1,0} = \bar{s} \circ \mathsf{T}(\widehat{S}_{1,-1})$;
for $t = 1, \ldots, k_{\text{out}}$ **do**
  for $k = 0, \ldots, k_{\text{in}} - 2$ **do**
    Sample a mini-batch $\mathcal{B}_{t,k+1}$ in $\{1, \ldots, n\}$ of size b, with or without replacement;
    $\mathsf{S}_{t,k+1} = \mathsf{S}_{t,k} + \bar{s}_{\mathcal{B}_{t,k+1}} \circ \mathsf{T}(\widehat{S}_{t,k}) - \bar{s}_{\mathcal{B}_{t,k+1}} \circ \mathsf{T}(\widehat{S}_{t,k-1})$;
    $\widehat{S}_{t,k+1} = \widehat{S}_{t,k} + \gamma_{t,k+1}\big(\mathsf{S}_{t,k+1} - \widehat{S}_{t,k}\big)$
  $\widehat{S}_{t+1,-1} = \widehat{S}_{t,k_{\text{in}}-1}$;
  $\mathsf{S}_{t+1,0} = \bar{s} \circ \mathsf{T}(\widehat{S}_{t+1,-1})$;
  $\widehat{S}_{t+1,0} = \widehat{S}_{t,k_{\text{in}}-1} + \gamma_{t,k_{\text{in}}}\big(\mathsf{S}_{t+1,0} - \widehat{S}_{t,k_{\text{in}}-1}\big)$

Algorithm 1: The `SPIDER-EM` algorithm.

We discuss the design considerations of the `SPIDER-EM` algorithm and provide insights on how it can accelerate convergence as follows.

Control Variate and Variance Reduction. We shall analyze SPIDER-EM as an SA scheme with control variate to reduce variance. While the description of SPIDER-EM algorithm in the above does not present the control variates explicitly, it is possible to re-interpret the inner loop (line 4–line 6) with a control variate defined, for $t \in \mathbb{N}_\star$ and $k \in \{0, \dots, k_{\rm in} - 2\}$, as

$$\begin{aligned} V_{t,k+1} &= V_{t,k} + \bar{s}_{\mathcal{B}_{t,k}} \circ \mathsf{T}(\widehat{S}_{t,k-1}) - \bar{s}_{\mathcal{B}_{t,k+1}} \circ \mathsf{T}(\widehat{S}_{t,k-1}) \\ &= \textstyle\sum_{j=0}^{k} \{\bar{s}_{\mathcal{B}_{t,j}} \circ \mathsf{T}(\widehat{S}_{t,j-1}) - \bar{s}_{\mathcal{B}_{t,j+1}} \circ \mathsf{T}(\widehat{S}_{t,j-1})\}, \end{aligned} \tag{12}$$

where $V_{t,0} = 0$ is reset at every outer iteration and, by convention, $\mathcal{B}_{t,0} \stackrel{\rm def}{=} \{1, \dots, n\}$. It is seen that line 6 can be rewritten as (see Lemma 3 in the supplementary material)

$$\widehat{S}_{t,k+1} = \widehat{S}_{t,k} + \gamma_{t,k+1}\big(\bar{s}_{\mathcal{B}_{t,k+1}} \circ \mathsf{T}(\widehat{S}_{t,k}) - \widehat{S}_{t,k} + V_{t,k+1}\big) \; . \tag{13}$$

Note that, by construction, the control variate $V_{t,k}$ is zero mean because, $\mathbb{E}[\bar{s}_{\mathcal{B}_{t,j}} \circ \mathsf{T}(\widehat{S}_{t,j-1})] = \mathbb{E}[\bar{s}_{\mathcal{B}_{t,j+1}} \circ \mathsf{T}(\widehat{S}_{t,j-1})] = \mathbb{E}[\bar{s} \circ \mathsf{T}(\widehat{S}_{t,j-1})]$. Eq. (12) shows how SPIDER-EM constructs a control variate by accumulating information – similar to SPIDER and SARAH in the gradient descent setting.

Comparing (12)-(13) to (9)-(10), the SPIDER-EM algorithm differs from sEM-vr only in the construction of the control variate. To obtain insights about their performance, let us denote the filtration as $\mathcal{F}_{t,k} \stackrel{\rm def}{=} \sigma(\widehat{S}_{\rm init}, \mathcal{B}_{1,1}, \dots, \mathcal{B}_{1,k_{\rm in}-1}, \dots, \mathcal{B}_{t,1}, \dots, \mathcal{B}_{t,k})$. Observe that the conditional variances (given $\mathcal{F}_{t,k}$) of $\widehat{S}_{t,k+1}$ of the sEM-VR and SPIDER-EM algorithms are:

$$\begin{aligned} \mathrm{Var}\left[\widehat{S}^{\texttt{sEM-vr}}_{t,k+1} | \mathcal{F}_{t,k}\right] &= \gamma^2_{t,k+1} \, \mathrm{Var}[\bar{s}_{\mathcal{B}_{t,k+1}} \circ \mathsf{T}(\widehat{S}_{t,k}) - \bar{s}_{\mathcal{B}_{t,k+1}} \circ \mathsf{T}(\widehat{S}_{t-1,k_{\rm in}-1}) | \mathcal{F}_{t,k}] \; , \\ \mathrm{Var}\left[\widehat{S}^{\texttt{SPIDER-EM}}_{t,k+1} | \mathcal{F}_{t,k}\right] &= \gamma^2_{t,k+1} \, \mathrm{Var}[\bar{s}_{\mathcal{B}_{t,k+1}} \circ \mathsf{T}(\widehat{S}_{t,k}) - \bar{s}_{\mathcal{B}_{t,k+1}} \circ \mathsf{T}(\widehat{S}_{t,k-1}) | \mathcal{F}_{t,k}] \; . \end{aligned}$$

As a comparison, the variance of $\widehat{S}_{(t-1)k_{\rm in}+k+1}$ for the Online EM is given by

$$\gamma^2_{(t-1)k_{\rm in}+k+1} \, \mathrm{Var}\left[\bar{s}_{\mathcal{B}_{(t-1)k_{\rm in}+k+1}} \circ \mathsf{T}(\widehat{S}_{(t-1)k_{\rm in}+k}) | \mathcal{F}^{\texttt{O-EM}}_{(t-1)k_{\rm in}+k}\right].$$

Here, $\mathcal{F}^{\texttt{O-EM}}_{\tau} \stackrel{\rm def}{=} \sigma(\widehat{S}_{\rm init}, \mathcal{B}_1, \dots, \mathcal{B}_\tau)$. In this sense, both sEM-vr and SPIDER-EM are variance-reduced versions of the Online EM. Additionally, SPIDER-EM and sEM-VR are designed to exploit two values $\widehat{S}_{t,k}, \widehat{S}_{t,k-1}$ and $\widehat{S}_{t,k}, \widehat{S}_{t-1,k_{\rm in}-1}$, respectively. The former thus takes the benefit of a stronger correlation between two successive values of $\{\widehat{S}_{t,k}, k \geq 1\}$ than between $\widehat{S}_{t,k}$ and $\widehat{S}_{t-1,k_{\rm in}-1}$ in the variance reduction step. As a result, SPIDER-EM should inherit a better rate of convergence – an intuition which is established will be Theorem 2.

Step Size and Memory Footprint. The SPIDER-EM algorithm is described with a positive step size sequence $\{\gamma_{t,k+1}, t \geq 1, k \geq 0\}$. Different strategies are allowed: (a) a constant step size $\gamma_{t,k+1} = \gamma$ for any $k \geq 0$, or (b) a random sequence. We focus on case (a) in the following, while we refer the readers to [11] for such a strategy in the gradient setting. Lastly, we observe that the SPIDER-EM algorithm has the same memory footprint requirement as the sEM-vr algorithm.

Convergence Analysis. Let (τ, ξ) be uniform r.v. on $\{1, \dots, k_{\rm out}\} \times \{0, \dots, k_{\rm in} - 1\}$, independent of the SPIDER-EM sequence $\{\widehat{S}_{t,k}, t = 1, \cdots, k_{\rm out}; k = -1, \cdots, k_{\rm in} - 1\}$. Our goal is to derive explicit upper bounds for $\mathbb{E}[\|h(\widehat{S}_{\tau,\xi-1})\|^2]$ for the SPIDER-EM sequence given by algorithm 1 with a constant step size ($\gamma_{t,k+1} = \gamma$ for any $t \geq 1$, $k \geq 0$). We strengthen the assumption H4 as follows:

H5. *(a) There exist* $0 < v_{\rm min} \leq v_{\rm max} < \infty$ *such that for all* $s \in \mathbb{R}^q$*, the spectrum of* $B(s)$ *is in* $[v_{\rm min}, v_{\rm max}]$*;* $B(s)$ *is defined in H4.*

(b) For any $i \in \{1, \dots, n\}$*, the map* $\bar{s}_i \circ \mathsf{T}$ *is globally Lipschitz on* $\mathbb{R}^q$ *with constant* L_i*.*

(c) The function $s \mapsto \nabla \mathrm{W}(s) = -B(s)h(s)$ *is globally Lipschitz on* $\mathbb{R}^q$ *with constant* $L_{\nabla \mathrm{W}}$*.*

From H5-(a) and Proposition 1, we have $\mathbb{E}\big[\|h(\widehat{S}_{\tau,\xi-1})\|^2\big] \geq v^{-2}_{\rm max} \mathbb{E}\big[\|\nabla \mathrm{W}(\widehat{S}_{\tau,\xi-1})\|^2\big]$ so that a control of $\mathbb{E}\big[\|h(\widehat{S}_{\tau,\xi-1})\|^2\big]$ provides a control of $\mathbb{E}\big[\|\nabla \mathrm{W}(\widehat{S}_{\tau,\xi-1})\|^2\big]$. The convergence result for SPIDER-EM is summarized below:

Theorem 2. *Assume H1, H2, H3, H4 and H5 and set* $L^2 \overset{\text{def}}{=} n^{-1}\sum_{i=1}^n L_i^2$. *Fix* $k_{\text{out}}, k_{\text{in}} \in \mathbb{N}_\star$, $\mathsf{b} \in \mathbb{N}_\star$ *and set* $\gamma_{t,k} \overset{\text{def}}{=} \alpha/L$ *for any* $t, k > 0$ *where* $\alpha \in (0, v_{\min}/\mu_\star(k_{\text{in}}, \mathsf{b}))$ *with*

$$\mu_\star(k_{\text{in}}, \mathsf{b}) \overset{\text{def}}{=} v_{\max}\sqrt{k_{\text{in}}/\mathsf{b}} + L_{\nabla \mathrm{W}}/(2L) \,. \tag{14}$$

The SPIDER-EM *sequence* $\{\widehat{S}_{t,k}, t \geq 1, k \geq 0\}$ *given by algorithm 1 satisfies*

$$\mathbb{E}\left[\|h(\widehat{S}_{\tau,\xi-1})\|^2\right] \leq \left(\frac{1}{k_{\text{in}}} + \frac{\alpha^2}{\mathsf{b}}\right) \frac{2L}{\alpha\{v_{\min} - \alpha\mu_\star(k_{\text{in}}, \mathsf{b})\}} \frac{1}{k_{\text{out}}} \left(\mathbb{E}[\mathrm{W}(\widehat{S}_{\text{init}})] - \min \mathrm{W}\right) .$$

Our analysis, whose detail can be found in the supplementary material, shares some similarities with the one in SPIDER-Boost [27]. Nevertheless, there are a number of differences because (a) SPIDER-EM algorithm recursion uses two spaces (the expectation space and the parameter space) which are connected by the maps $\bar{s}$ and T; (b) SPIDER-EM is not a gradient algorithm in the expectation space, but an SA scheme to obtain a root for h; (c) there is a Lyapunov function $\mathrm{W}(s)$ where $\nabla\mathrm{W}(s) \neq -h(s)$, but which satisfies $\langle \nabla\mathrm{W}(s), h(s)\rangle \leq -v_{\min}\|h(s)\|^2$. In addition, in relation to the above points, our analysis took insights from [16, 17] to analyze SPIDER-EM as a biased SA scheme. Our challenge lies in carefully controlling the bias/variance of the SPIDER estimator employed, which is not reported in the prior literature.

Proof Sketch. While we shall omit the proof details, an outline of the proof is provided. Set $H_{t,k+1} \overset{\text{def}}{=} \gamma_{t,k+1}^{-1}(\widehat{S}_{t,k+1} - \widehat{S}_{t,k})$. A key property is the following descent condition for the Lyapunov function W. There exist positive sequences $\Lambda_{t,k}, \beta_{t,k}$ such that for any $t \geq 1$, $k \geq 0$,

$$\mathrm{W}(\widehat{S}_{t,k+1}) \leq \mathrm{W}(\widehat{S}_{t,k}) - \Lambda_{t,k+1}\|H_{t,k+1}\|^2 + \gamma_{t,k+1}\frac{v_{\max}^2}{2\beta_{t,k+1}^2}\|H_{t,k+1} - h(\widehat{S}_{t,k})\|^2 \,.$$

It holds for any $t \geq 1$ and $0 \leq k \leq k_{\text{in}} - 2$,

$$\mathbb{E}\left[\|H_{t,k+1} - h(\widehat{S}_{t,k})\|^2 | \mathcal{F}_{t-1,k_{\text{in}}-1}\right] \leq \frac{L^2}{\mathsf{b}}\sum_{j=0}^k \gamma_{t,j}^2 \mathbb{E}\left[\|H_{t,j}\|^2 | \mathcal{F}_{t-1,k_{\text{in}}-1}\right] \,. \tag{15}$$

The above conditions can be combined to yield

$$\sum_{t=1}^{k_{\text{out}}}\sum_{k=0}^{k_{\text{in}}-1} A_{t,k}\mathbb{E}\left[\|H_{t,k}\|^2\right] \leq \mathbb{E}\left[\mathrm{W}(\widehat{S}_{\text{init}})\right] - \min \mathrm{W}$$

where the $A_{t,k}$'s are positive. Dividing both sides of the inequality by $\sum_{t=1}^{k_{\text{out}}}\sum_{k=0}^{k_{\text{in}}-1} A_{t,k}$ leads to a bound on $\mathbb{E}[\|H_\Xi\|^2]$ for some r.v. Ξ on $\{1, \ldots, k_{\text{out}}\} \times \{0, \ldots, k_{\text{in}} - 1\}$. For the concerned case when $\gamma_{t,k} = \gamma$, we have $A_{t,k} = A$ and $\Xi = (\tau, \xi)$ is the uniform distribution, thus the convergence rate for $\mathbb{E}[\|H_{\tau,\xi}\|^2]$ is $\mathcal{O}(1/k_{\text{in}}k_{\text{out}})$. Lastly, we obtain a bound for the mean field $\|h(\widehat{S}_{\tau,\xi-1})\|^2$ using the standard inequality $(a+b)^2 \leq 2a^2 + 2b^2$ and (15) again.

Choice of $k_{\text{in}}, \mathsf{b}, k_{\text{out}}$ and Complexity Bounds. The maximum of $\alpha\{v_{\min} - \alpha\mu_\star(k_{\text{in}}, \mathsf{b})\}$ on $(0, v_{\min}/\mu_\star(k_{\text{in}}, \mathsf{b}))$ is $\alpha_\star(k_{\text{in}}, \mathsf{b}) \overset{\text{def}}{=} v_{\min}/\{2\mu_\star(k_{\text{in}}, \mathsf{b})\}$ which yields $\gamma = v_{\min}/\{2\mu_\star(k_{\text{in}}, \mathsf{b})L\}$ and the upper bound

$$\mathbb{E}\left[\|h(\widehat{S}_{\tau,\xi-1})\|^2\right] \leq \left(\frac{\mu_\star(k_{\text{in}}, \mathsf{b})}{v_{\min}^2} + \frac{k_{\text{in}}}{4\mu_\star(k_{\text{in}}, \mathsf{b})\mathsf{b}}\right) \frac{8L}{k_{\text{in}}k_{\text{out}}} (\mathbb{E}[\mathrm{W}(\widehat{S}_{\text{init}})] - \min \mathrm{W}) \,.$$

The number of parameter updates is $1 + k_{\text{out}} + k_{\text{in}}k_{\text{out}}$. The number of per-sample conditional expectation computations is $n + k_{\text{out}}n + 2\mathsf{b}k_{\text{in}}k_{\text{out}}$. Assume that n and $\epsilon > 0$ are given. Set for simplicity $\mathsf{b} = k_{\text{in}} = \lceil\sqrt{n}\rceil$ which means that the number of per-sample conditional expectations evaluations in the inner loop is equal to n, i.e., is an epoch (see subsection 9.3 for a discussion on other strategies). With this choice, we get $\mu_\star(k_{\text{in}}, \mathsf{b}) = m_\star \overset{\text{def}}{=} v_{\max} + L_{\nabla \mathrm{W}}/(2L)$. Taking

$$k_{\text{out}} \geq \left(\frac{m_\star}{v_{\min}^2} + \frac{1}{4m_\star}\right) \frac{8L}{\sqrt{n}\epsilon} (\mathbb{E}[\mathrm{W}(\widehat{S}_{\text{init}})] - \min \mathrm{W}) \,,$$

then we have $\mathbb{E}[\|h(\widehat{S}_{\tau,\xi-1})\|^2] \leq \epsilon$. With these choices of $k_{\text{in}}, k_{\text{out}}, \mathsf{b}$, the complexity in terms of the number of per-sample conditional expectations evaluations $\bar{s}_i$ is $K_{\text{CE}}(n, \epsilon) = n + \sqrt{n}L\mathcal{O}(\epsilon^{-1})$. The number of parameter updates is $K_{\text{Opt}}(n, \epsilon) = \mathcal{O}(\epsilon^{-1})$. Note that the step size is chosen to be $\gamma = \alpha_\star(k_{\text{in}}, \mathsf{b})/L$, which is independent of the targeted accuracy ϵ.

Linear convergence rate. In section 10, we provide a modification of SPIDER-EM which exhibits a linear convergence rate when W satisfies a Polyak-Lojasiewicz inequality. Note that the latter condition (or its variants) has been used in a few recent works, e.g., [1, 7].

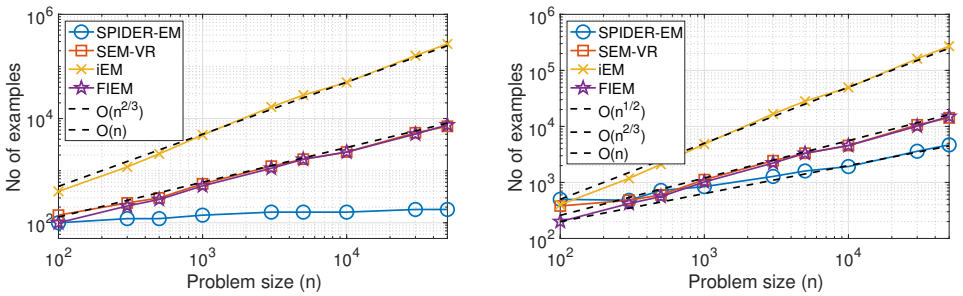

Figure 1: [Left] Median estimated number of parameter updates $K_{\mathrm{Opt}}(n,\epsilon)$ needed to reach an accuracy of 2.5×10^{-5} [Right] Median estimated number of per-sample conditional expectations $K_{\mathrm{CE}}(n,\epsilon) - n$ needed to reach an accuracy of 2.5×10^{-5}. The median is taken from a Monte-Carlo simulation among 50 trials.

4 Numerical illustration

Synthetic Data. We evaluate the efficiency of `SPIDER-EM` against the problem size. We generate a synthetic dataset with n observations from a scalar two-components Gaussian mixture model (GMM) with $0.2\mathcal{N}(0.5,1) + 0.8\mathcal{N}(-0.5,1)$. The variances and the weights are assumed known. We fit the means μ_1, μ_2 of a GMM to the observed data. For `SPIDER-EM`, we set $\mathsf{b} = \lceil \sqrt{n}/20 \rceil$, $k_{\mathrm{in}} = \lceil n/\mathsf{b} \rceil$ and a fixed step size $\gamma_k = 0.01$. We define $\tau_{\mathrm{emp}} = t_{\mathrm{emp}} k_{\mathrm{in}} + k_{\mathrm{emp}}$ as the total number of updates of $\widehat{S}_k$ evaluated, such that t_{emp}, k_{emp} are the indices of outer, inner iteration, respectively. To estimate $K_{\mathrm{Opt}}(n,\epsilon)$ and $K_{\mathrm{CE}}(n,\epsilon)$, we run the `SPIDER-EM` algorithm until the first iteration τ_{emp} when the solution satisfies $\|h(\widehat{S}_{t_{\mathrm{emp}},k_{\mathrm{emp}}})\|^2 \leq \epsilon = 2.5 \times 10^{-5}$. We take the median of τ_{emp} over 50 runs to give an estimate of $K_{\mathrm{Opt}}(n,\epsilon)$; similarly, we take the median of $nt_{\mathrm{emp}} + 2\mathsf{b}\tau_{\mathrm{emp}}$ to give an estimate of $K_{\mathrm{CE}}(n,\epsilon)$. Note that the conditional expectations computed during the initialization step are ignored.

Figure 1 compares `SPIDER-EM` to the state-of-the-art incremental EM algorithms for different settings of n. The results illustrate that the empirical performance of `SPIDER-EM` agrees with the theoretical analysis. In particular, we observe that for `SPIDER-EM`, the estimated $K_{\mathrm{Opt}}(n,\epsilon)$ is independent of the problem size n while $K_{\mathrm{CE}}(n,\epsilon) - n$ grows at the rate of $\sqrt{n}$.

MNIST Dataset. We perform experiment on the MNIST dataset to illustrate the effectiveness of `SPIDER-EM` on real data; this example is taken from [23, Section 5]. The dataset consists of $n = 6 \times 10^4$ images of handwritten digits, each with 784 pixels. We pre-process the dataset as follows. First, we eliminate the uninformative pixels (67 pixels are always zero) across all images to obtain a dense representation with $d_{\mathrm{dense}} = 717$ pixels per image. Second, we apply principal component analysis (PCA) to further reduce the data dimension. We keep the $d_{\mathrm{PC}} = 20$ principal components (PCs) of each observation.

We estimate a multivariate GMM model with $g = 12$ components. Unlike in the previous experiment, here the parameter θ collects the mixture's weights $\{\alpha_\ell, 1 \leq \ell \leq g\}$, the expectations of each component and a pulled full covariance matrix. `SPIDER-EM` is compared to `iEM` [21], `Online EM` [6], `FIEM` [18], and `sEM-vr` [7]. Details on the multivariate Gaussian mixture model are given in the supplementary material, section 11, where we give technical conditions required to verify the assumptions of Theorem 2.

In Figure 2, we display the sequence of parameter estimates $\{\theta_\tau\}$, the objective function $\{-\mathrm{W}(\widehat{S}_\tau)\}$ and the squared norm of the mean field $\{\|h(\widehat{S}_\tau)\|^2\}$. Figure 3 gives insights on the distribution of $\|h(\widehat{S}_{t,k})\|^2$ along `SPIDER-EM` paths. The mini-batches $\{\mathcal{B}_\tau\}_\tau$ are independent, and sampled at random in $\{1,\ldots,n\}$ with replacement. For a fair comparison, we use the same seed to sample the minibatches $\{\mathcal{B}_k\}$; another seed is used for `FIEM` which requires a second sequence of minibatches $\{\overline{\mathcal{B}}_\tau\}_\tau$. The minibatch size is set to be $\mathsf{b} = 100$ and the stepsize $\gamma_\tau = 5 \times 10^{-3}$ except for `iEM` where $\gamma_\tau = 1$. The same initial value $\widehat{S}_{\mathrm{init}}$ is used for all experiments. We have implemented the procedure of [19] in order to obtain the initialization θ_{init} and then we set $\widehat{S}_{\mathrm{init}} \overset{\mathrm{def}}{=} \bar{s}(\theta_{\mathrm{init}})$ ($-\mathrm{W}(\widehat{S}_{\mathrm{init}}) =$

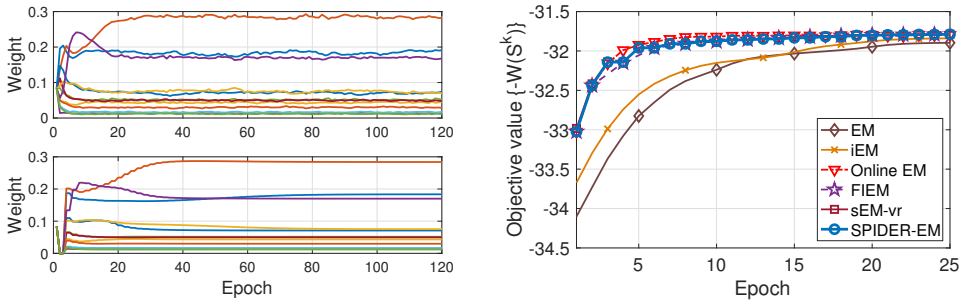

Figure 2: [Left] Evolution of the estimates of the weights α_ℓ for $\ell = 1, \ldots, g$ by `Online EM` (top) and `SPIDER-EM` (bottom) vs the number of epochs. [Right] Evolution of the objective function $-\mathrm{W}(\widehat{S}_k)$ vs the number of epochs.

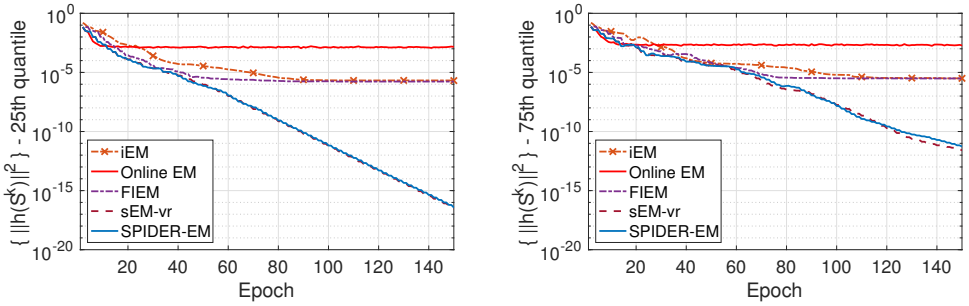

Figure 3: [Left] Quantile 0.25 and [Right] quantile 0.75 of the distribution of $\|h(\widehat{S}_{t,-1})\|^2$ vs the number of epochs t; the quantiles are estimated from 40 independent samples of this distribution.

-58.3). The plots illustrate that `SPIDER-EM` reduces the variability of `Online EM` and compares favorably to `iEM` and `FIEM`. Additional details and results are given in the Supplementary material.

5 Conclusions

We have introduced the `SPIDER-EM` algorithm for large-scale inference. The algorithm offers low memory footprint and improved complexity bounds compared to the state-of-the-art, which is verified by theoretical analysis and numerical experiments.

Broader Impact This work does not present any foreseeable societal consequence.

Acknowledgments and Disclosure of Funding

The work of G. Fort is partially supported by the *Fondation Simone et Cino del Duca* under the project OpSiMorE. The work of E. Moulines is partially supported by ANR-19-CHIA-0002-01 / chaire SCAI. It was partially prepared within the framework of the HSE University Basic Research Program. The work of H.-T. Wai is partially supported by the CUHK Direct Grant #4055113.

References

[1] S. Balakrishnan, M. J. Wainwright, and B. Yu. Statistical guarantees for the EM algorithm: From population to sample-based analysis. *Ann. Statist.*, 45(1):77–120, 2017.

[2] A. Benveniste, M. Métivier, and P. Priouret. *Adaptive Algorithms and Stochastic Approximations*. Springer Verlag, 1990.

[3] V. S. Borkar. *Stochastic approximation*. Cambridge University Press, Cambridge; Hindustan Book Agency, New Delhi, 2008. A dynamical systems viewpoint.

[4] L. Bottou and Y. Le Cun. Large scale online learning. In S. Thrun, L. K. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*, pages 217–224. MIT Press, 2004.

[5] P. Bühlmann, P. Drineas, M. Kane, and M. van der Laan. *Handbook of Big Data*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. CRC Press, 2016. ISBN 9781482249088.

[6] O. Cappé and E. Moulines. On-line expectation-maximization algorithm for latent data models. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 71(3):593–613, 2009.

[7] J. Chen, J. Zhu, Y. Teh, and T. Zhang. Stochastic expectation maximization with variance reduction. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 7967–7977. 2018.

[8] A. Defazio, F. Bach, and S. Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 1646–1654. Curran Associates, Inc., 2014.

[9] B. Delyon, M. Lavielle, and E. Moulines. Convergence of a stochastic approximation version of the EM algorithm. *Ann. Statist.*, 27(1):94–128, 1999. ISSN 0090-5364. doi: 10.1214/aos/1018031103. URL https://doi.org/10.1214/aos/1018031103.

[10] A. Dempster, N. Laird, and D. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *J. Roy. Stat. Soc. B Met.*, 39(1):1–38, 1977.

[11] C. Fang, C. Li, Z. Lin, and T. Zhang. SPIDER: Near-Optimal Non-Convex Optimization via Stochastic Path-Integrated Differential Estimator. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 689–699. Curran Associates, Inc., 2018.

[12] G. Fort, P. Gach, and E. Moulines. Fast Incremental Expectation Maximization for non-convex optimization: non asymptotic convergence bounds. Technical report, HAL-02617725v1, 2020.

[13] S. Ghadimi and G. Lan. Stochastic First- and Zeroth-Order Methods for Nonconvex Stochastic Programming. *SIAM J. Optimiz.*, 23(4):2341–2368, 2013.

[14] W. Härdle, H. H.-S. Lu, and X. Shen. *Handbook of big data analytics*. Springer, 2018.

[15] R. Johnson and T. Zhang. Accelerating Stochastic Gradient Descent using Predictive Variance Reduction. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 315–323. Curran Associates, Inc., 2013.

[16] B. Karimi, M. Lavielle, and E. Moulines. On the Convergence Properties of the Mini-Batch EM and MCEM Algorithms. Technical report, hal-02334485, 2019.

[17] B. Karimi, B. Miasojedow, E. Moulines, and H.-T. Wai. Non-asymptotic Analysis of Biased Stochastic Approximation Scheme. In *COLT*, 2019.

[18] B. Karimi, H.-T. Wai, E. Moulines, and M. Lavielle. On the Global Convergence of (Fast) Incremental Expectation Maximization Methods. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 2837–2847. Curran Associates, Inc., 2019.

[19] W. Kwedlo. A new random approach for initialization of the multiple restart EM algorithm for Gaussian model-based clustering. *Pattern Anal. Applic.*, 18:757–770, 2015.

[20] G. McLachlan and T. Krishnan. *The EM algorithm and extensions*. Wiley series in probability and statistics. Wiley, 2008.

[21] R. M. Neal and G. E. Hinton. A View of the EM Algorithm that Justifies Incremental, Sparse, and other Variants. In M. I. Jordan, editor, *Learning in Graphical Models*, pages 355–368. Springer Netherlands, Dordrecht, 1998.

[22] S. K. Ng and G. J. McLachlan. On the choice of the number of blocks with the incremental EM algorithm for the fitting of normal mixtures. *Stat. Comput.*, 13(1):45–55, 2003.

[23] H. Nguyen, F. Forbes, and G. McLachlan. Mini-batch learning of exponential family finite mixture models. *Stat. Comput.*, 2020.

[24] L. M. Nguyen, J. Liu, K. Scheinberg, and M. Takáč. Sarah: A novel method for machine learning problems using stochastic recursive gradient. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 2613–2621. JMLR.org, 2017.

[25] S. Reddi, S. Sra, B. Póczos, and A. Smola. Fast Incremental Method for Smooth Nonconvex Optimization. In *2016 IEEE 55th Conference on Decision and Control (CDC)*, pages 1971–1977, 2016.

[26] H. Robbins and S. Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[27] Z. Wang, K. Ji, Y. Zhou, Y. Liang, and V. Tarokh. SpiderBoost and Momentum: Faster Stochastic Variance Reduction Algorithms. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 2406–2416. 2019.

Supplementary materials for "A Stochastic Path-Integrated Differential EstimatoR Expectation Maximization Algorithm"

Gersende Fort
Institut de Mathématiques de Toulouse
Université de Toulouse; CNRS
UPS, F-31062 Toulouse Cedex 9, France
gersende.fort@math.univ-toulouse.fr

Eric Moulines
Centre de Mathématiques Appliquées
Ecole Polytechnique, France
CS Departement
HSE University, Russian Federation
eric.moulines@polytechnique.edu

Hoi-To Wai
Department of SEEM
The Chinese University of Hong Kong
Shatin, Hong Kong
htwai@cuhk.edu.hk

Notations. For two vectors $a, b \in \mathbb{R}^r$, $\langle a, b \rangle$ denotes the usual Euclidean product and $\|a\|$ the associated norm. By convention, vectors are column vectors. For a vector x with components $(x_1, \ldots, x_r)$, $x_{i:j}$ denotes the sub-vector with components $(x_i, x_{i+1}, \ldots, x_{j-1}, x_j)$.

For two matrices $A \in \mathbb{R}^{r_1 \times r_2}$ and $B \in \mathbb{R}^{r_3 \times r_4}$, $A \otimes B$ denotes the Kronecker product. I_r is the $r \times r$ identity matrix. A^T is the transpose of A.

6 Complexity of incremental EM-based methods for smooth non-convex finite sum optimization

We first compare the complexities of the incremental EM based methods using the following table which summarizes the state-of-the-art results.

algorithm	γ	K_{Opt}	K_{CE}	Optimal K_{CE}
EM [10]	-	$1 + k_{\max}$	$n + n k_{\max}$	N/A
online-EM [6]	decaying; $\mathcal{O}(L^{-1} k^{-1/2})$	$1 + k_{\max}$	$n + \mathsf{b} k_{\max}$	ϵ^{-2}
iEM [21]	1	$1 + k_{\max}$	$n + \mathsf{b} k_{\max}$	$\epsilon^{-1} n$
sEM-vr [7, 18]	$O(L^{-1} n^{-2/3})$	$1 + k_{\mathrm{in}} k_{\mathrm{out}}$	$n(1 + k_{\mathrm{out}}) + (\mathsf{b} k_{\mathrm{in}} + n) k_{\mathrm{out}}$	$\epsilon^{-1} n^{2/3}$
FIEM [18]	$O(L^{-1} n^{-2/3})$	$1 + k_{\max}$	$n + 2\mathsf{b} k_{\max}$	$\epsilon^{-1} n^{2/3}$
FIEM [12]	$O(L^{-1} n^{-1/3} k_{\max}^{-1/3})$	$1 + k_{\max}$	$n + 2\mathsf{b} k_{\max}$	$\epsilon^{-3/2} \sqrt{n}$
SPIDER-EM	$O(L^{-1})$	$1 + k_{\mathrm{in}} k_{\mathrm{out}}$	$n + k_{\mathrm{out}} n + 2\mathsf{b} k_{\mathrm{in}} k_{\mathrm{out}}$	$\epsilon^{-1} \sqrt{n}$

Table 1: Comparison between different EM-based algorithms for smooth non convex finite sum optimization. Except sEM-vr and SPIDER-EM which have nested loops (k_{out} is the maximal number of outer loops and k_{in} is the number of inner loops per outer loop), $k_{\max}$ is the maximal number of iterations. The last column is the optimal complexity to reach an ϵ-approximate stationary point.

Next, we provide the psuedo-codes of several existing incremental EM-based algorithms, following the notations defined in the main paper.

Data: $k_{\max} \in \mathbb{N}_\star$, $\widehat{S}_{\text{init}} \in \mathbb{R}^q$
Result: The EM sequence: $\widehat{S}_k, k = 0, \ldots, k_{\max}$

$\widehat{S}_0 = \bar{s} \circ \mathsf{T}(\widehat{S}_{\text{init}})$;
for $k = 0, \ldots, k_{\max} - 1$ **do**
  $\widehat{S}_{k+1} = \bar{s} \circ \mathsf{T}(\widehat{S}_k)$

Algorithm 2: The EM algorithm in the expectation space.

Data: $k_{\max} \in \mathbb{N}_\star$, $\widehat{S}_{\text{init}} \in \mathbb{R}^q$, $\gamma_k \in (0, \infty)$ for $k = 1, \ldots, k_{\max}$
Result: The SA sequence: $\widehat{S}_k, k = 0, \ldots, k_{\max}$

$\widehat{S}_0 = \bar{s} \circ \mathsf{T}(\widehat{S}_{\text{init}})$;
for $k = 0, \ldots, k_{\max} - 1$ **do**
  Sample a mini-batch $\mathcal{B}_{k+1}$ in $\{1, \ldots, n\}$ of size b, with replacement ;
  $\widehat{S}_{k+1} = \widehat{S}_k + \gamma_{k+1} \left(\bar{s}_{\mathcal{B}_{k+1}} \circ \mathsf{T}(\widehat{S}_k) - \widehat{S}_k \right)$.

Algorithm 3: The Online EM algorithm.

Data: $k_{\max} \in \mathbb{N}_\star$, $\widehat{S}_{\text{init}} \in \mathbb{R}^q$, $\gamma_k \in (0, \infty)$ for $k = 1, \ldots, k_{\max}$
Result: The iEM sequence: $\widehat{S}_k, k = 0, \ldots, k_{\max}$

$\mathsf{S}_{0,i} = \bar{s}_i \circ \mathsf{T}(\widehat{S}_{\text{init}})$ for all $i = 1, \ldots, n$;
$\widehat{S}_0 = \widetilde{S}_0 = n^{-1} \sum_{i=1}^n \mathsf{S}_{0,i}$;
for $k = 0, \ldots, k_{\max} - 1$ **do**
  Sample a mini-batch $\mathcal{B}_{k+1}$ in $\{1, \ldots, n\}$ of size b, with replacement ;
  $\mathsf{S}_{k+1,i} = \mathsf{S}_{k,i}$ for $i \notin \mathcal{B}_{k+1}$;
  $\mathsf{S}_{k+1,i} = \bar{s}_i \circ \mathsf{T}(\widehat{S}_k)$ for $i \in \mathcal{B}_{k+1}$;
  $\widetilde{S}_{k+1} = \widetilde{S}_k + n^{-1} \sum_{i \in \mathcal{B}_{k+1}} (\mathsf{S}_{k+1,i} - \mathsf{S}_{k,i})$;
  $\widehat{S}_{k+1} = \widehat{S}_k + \gamma_{k+1}(\widetilde{S}_{k+1} - \widehat{S}^k)$

Algorithm 4: The Incremental EM (iEM) algorithm.

Data: $k_{\max} \in \mathbb{N}_\star$, $\widehat{S}_{\text{init}} \in \mathbb{R}^q$, $\gamma_k \in (0, \infty)$ for $k = 1, \ldots, k_{\max}$
Result: The FIEM sequence: $\widehat{S}_k, k = 0, \ldots, k_{\max}$

$\mathsf{S}_{0,i} = \bar{s}_i \circ \mathsf{T}(\widehat{S}_{\text{init}})$ for all $i = 1, \ldots, n$;
$\widehat{S}_0 = \widetilde{S}_0 = n^{-1} \sum_{i=1}^n \mathsf{S}_{0,i}$;
for $k = 0, \ldots, k_{\max} - 1$ **do**
  Sample a mini-batch $\mathcal{B}_{k+1}$ in $\{1, \ldots, n\}$ of size b, with replacement ;
  $\mathsf{S}_{k+1,i} = \mathsf{S}_{k,i}$ for $i \notin \mathcal{B}_{k+1}$;
  $\mathsf{S}_{k+1,i} = \bar{s}_i \circ \mathsf{T}(\widehat{S}_k)$ for $i \in \mathcal{B}_{k+1}$;
  $\widetilde{S}_{k+1} = \widetilde{S}_k + n^{-1} \sum_{i \in \mathcal{B}_{k+1}} (\mathsf{S}_{k+1,i} - \mathsf{S}_{k,i})$;
  Sample a mini-batch $\mathcal{B}'_{k+1}$ in $\{1, \ldots, n\}$ of size b, with replacement ;
  $V_{k+1} = \widetilde{S}_{k+1} - \mathsf{b}^{-1} \sum_{i \in \mathcal{B}'_{k+1}} \mathsf{S}_{k+1,i}$;
  $\widehat{S}_{k+1} = \widehat{S}_k + \gamma_{k+1}(\bar{s}_{\mathcal{B}'_{k+1}} \circ \mathsf{T}(\widehat{S}_k) - \widehat{S}_k + V_{k+1})$

Algorithm 5: The Fast Incremental EM (FIEM) algorithm.

Data: $k_{\text{in}} \in \mathbb{N}_\star$, $k_{\text{out}} \in \mathbb{N}_\star$, $\widehat{S}_{\text{init}} \in \mathbb{R}^q$, $\gamma_{t,k} \in (0,\infty)$ for $t \geq 1, k \geq 1$
Result: The sEM-vr sequence: $\widehat{S}_{t,k}, t = 1, \ldots, k_{\text{out}}$ and $k = 0, \ldots, k_{\text{in}} - 1$

$\mathsf{S}_{1,0} = \bar{s} \circ \mathsf{T}(\widehat{S}_{\text{init}})$;
$\widehat{S}_{1,0} = \widehat{S}_{\text{init}}$;
for $t = 1, \ldots, k_{\text{out}}$ **do**
  for $k = 0, \ldots, k_{\text{in}} - 2$ **do**
    Sample a mini-batch $\mathcal{B}_{t,k+1}$ in $\{1, \ldots, n\}$ of size b, with replacement ;
    $V_{t,k+1} = \mathsf{S}_{t,0} - \bar{s}_{\mathcal{B}_{t,k+1}} \circ \mathsf{T}(\widehat{S}_{t-1,k_{\text{in}}-1})$;
    $\widehat{S}_{t,k+1} = \widehat{S}_{t,k} + \gamma_{t,k+1} \left(\bar{s}_{\mathcal{B}_{t,k+1}} \circ \mathsf{T}(\widehat{S}_{t,k}) - \widehat{S}_{t,k} + V_{t,k+1} \right)$
  $\mathsf{S}_{t+1,0} = \bar{s} \circ \mathsf{T}(\widehat{S}_{t,k_{\text{in}}-1})$;
  $\widehat{S}_{t+1,0} = \widehat{S}_{t,k_{\text{in}}-1} + \gamma_{t,k_{\text{in}}} \left(\mathsf{S}_{t+1,0} - \widehat{S}_{t,k_{\text{in}}-1} \right)$

Algorithm 6: The sEM-vr algorithm.

7 An equivalent definition of the SPIDER-EM algorithm

Using Lemma 3 below this page, we deduce that SPIDER-EM can be equivalently described by the following algorithm 7.

Data: $k_{\text{in}} \in \mathbb{N}_\star$, $k_{\text{out}} \in \mathbb{N}_\star$, $\widehat{S}_{\text{init}} \in \mathbb{R}^q$, a positive sequence $\{\gamma_{t,k}, t, k \geq 1\}$.
Result: The SPIDER-EM sequence: $\widehat{S}_{t,k}$, $t = 1, \ldots, k_{\text{out}}$, $k = 0, \ldots, k_{\text{in}} - 1$

$\widehat{S}_{1,-1} = \widehat{S}_{\text{init}}$;
$\widetilde{S}_{1,0} = \bar{s} \circ \mathsf{T}(\widehat{S}_{\text{init}})$;
for $t = 1, \ldots, k_{\text{out}}$ **do**
  $V_{t,0} = 0$;
  for $k = 0, \ldots, k_{\text{in}} - 2$ **do**
    Sample a mini-batch $\mathcal{B}_{t,k+1}$ in $\{1, \ldots, n\}$ of size b, with or without replacement ;
    $V_{t,k+1} = V_{t,k} + \widetilde{S}_{t,k} - \bar{s}_{\mathcal{B}_{t,k+1}} \circ \mathsf{T}(\widehat{S}_{t,k-1})$;
    $\widetilde{S}_{t,k+1} = \bar{s}_{\mathcal{B}_{t,k+1}} \circ \mathsf{T}(\widehat{S}_{t,k})$;
    $\widehat{S}_{t,k+1} = \widehat{S}_{t,k} + \gamma_{t,k+1} \left(\widetilde{S}_{t,k+1} - \widehat{S}_{t,k} + V_{t,k+1} \right)$
  $\widetilde{S}_{t+1,0} = \bar{s} \circ \mathsf{T}(\widehat{S}_{t,k_{\text{in}}-1})$;
  $\widehat{S}_{t+1,0} = \widehat{S}_{t,k_{\text{in}}-1} + \gamma_{t,k_{\text{in}}} \left(\widetilde{S}_{t+1,0} - \widehat{S}_{t,k_{\text{in}}-1} \right)$

Algorithm 7: The SPIDER-EM algorithm (equivalent description)

Lemma 3. *Let $\{\gamma_k, k \geq 1\}$ be a positive deterministic sequence and $\{\mathcal{B}_k, t, k \geq 1\}$ be a family of mini-batches sampled from $\{1, \ldots, n\}$. Fix $\widehat{S}_{-1}, \widehat{S}_0$ and S_0. Define for $k = 0, \cdots, k_{\text{in}} - 2$*

$$\begin{cases} \mathsf{S}_{k+1} \overset{\text{def}}{=} \mathsf{S}_k + \bar{s}_{\mathcal{B}_{k+1}} \circ \mathsf{T}(\widehat{S}_k) - \bar{s}_{\mathcal{B}_{k+1}} \circ \mathsf{T}(\widehat{S}_{k-1}) \,, \\ \widehat{S}_{k+1} \overset{\text{def}}{=} \widehat{S}_k + \gamma_{k+1} \left(\mathsf{S}_{k+1} - \widehat{S}_k \right) \,. \end{cases}$$

Set $\widetilde{S}_{-1} \overset{\text{def}}{=} \widehat{S}_{-1}$, $\widetilde{S}_0 \overset{\text{def}}{=} \widehat{S}_0$, $V_0 \overset{\text{def}}{=} 0$ and define for $k = 0, \ldots, k_{\text{in}} - 2$,

$$\begin{cases} V_{k+1} \overset{\text{def}}{=} V_k + \bar{s}_{\mathcal{B}_k} \circ \mathsf{T}(\widetilde{S}_{k-1}) - \bar{s}_{\mathcal{B}_{k+1}} \circ \mathsf{T}(\widetilde{S}_{k-1}) \,, \\ \widetilde{S}_{k+1} \overset{\text{def}}{=} \widetilde{S}_k + \gamma_{k+1} \left(\bar{s}_{\mathcal{B}_{k+1}} \circ \mathsf{T}(\widetilde{S}_k) - \widetilde{S}_k + V_{k+1} \right) \,; \end{cases}$$

by convention, set $\bar{s}_{\mathcal{B}_0} \circ \mathsf{T}(\widetilde{S}_{-1}) = \mathsf{S}_0$.

Then for any $k = -1, \ldots, k_{\text{in}} - 1$, $\widetilde{S}_k = \widehat{S}_k$.

Proof. We prove by induction that for any $k \geq 1$, $V_k = \mathsf{S}_k - \bar{s}_{\mathcal{B}_k} \circ \mathsf{T}(\widehat{S}_{k-1})$ and $\widetilde{S}_k = \widehat{S}_k$. We have by definition of V_0, $\bar{s}_{\mathcal{B}_0} \circ \mathsf{T}(\widetilde{S}_{-1})$, $\widetilde{S}_{-1}$ and S_1,

$$V_1 = \mathsf{S}_0 - \bar{s}_{\mathcal{B}_1} \circ \mathsf{T}(\widetilde{S}_{-1}) = \mathsf{S}_0 - \bar{s}_{\mathcal{B}_1} \circ \mathsf{T}(\widehat{S}_{-1}) = \mathsf{S}_1 - \bar{s}_{\mathcal{B}_1} \circ \mathsf{T}(\widehat{S}_0) \; .$$

In addition, by definition of $\widetilde{S}_0$, $\widetilde{S}_1$ and V_1, we have

$$\widetilde{S}_1 = \widehat{S}_0 + \gamma_1 \left(\bar{s}_{\mathcal{B}_1} \circ \mathsf{T}(\widehat{S}_0) - \widehat{S}_0 + \mathsf{S}_1 - \bar{s}_{\mathcal{B}_1} \circ \mathsf{T}(\widehat{S}_{-0}) \right) \; .$$

Assume that the property holds for any $0 \leq j \leq k$. Then, by definition of V_{k+1}, the induction assumption on V_k and the definition of S_{k+1}, it holds

$$\begin{aligned} V_{k+1} &= V_k + \bar{s}_{\mathcal{B}_k} \circ \mathsf{T}(\widetilde{S}_{k-1}) - \bar{s}_{\mathcal{B}_{k+1}} \circ \mathsf{T}(\widetilde{S}_{k-1}) \\ &= \mathsf{S}_k - \bar{s}_{\mathcal{B}_{k+1}} \circ \mathsf{T}(\widetilde{S}_{k-1}) = \mathsf{S}_{k+1} - \bar{s}_{\mathcal{B}_{k+1}} \circ \mathsf{T}(\widetilde{S}_k) \; . \end{aligned}$$

This concludes the induction for the property on $\{V_k, k \geq 0\}$. In addition, by the induction assumption on $\widetilde{S}_k$, the definition of V_{k+1}, the induction assumption on V_k and the definition of S_{k+1}, we have

$$\begin{aligned} \widetilde{S}_{k+1} &= \widehat{S}_k + \gamma_{k+1} \left(\bar{s}_{\mathcal{B}_{k+1}} \circ \mathsf{T}(\widehat{S}_k) - \widehat{S}_k + V_k + \bar{s}_{\mathcal{B}_k} \circ \mathsf{T}(\widehat{S}_{k-1}) - \bar{s}_{\mathcal{B}_{k+1}} \circ \mathsf{T}(\widehat{S}_{k-1}) \right) \\ &= \widehat{S}_k + \gamma_{k+1} \left(\bar{s}_{\mathcal{B}_{k+1}} \circ \mathsf{T}(\widehat{S}_k) - \widehat{S}_k + \mathsf{S}_k - \bar{s}_{\mathcal{B}_{k+1}} \circ \mathsf{T}(\widehat{S}_{k-1}) \right) \\ &= \widehat{S}_k + \gamma_{k+1} \left(\mathsf{S}_{k+1} - \widehat{S}_k \right) = \widehat{S}_{k+1} \; . \end{aligned}$$

This concludes the proof. □

8 General convergence results

The purpose of this section is to show the general convergence results of a `SPIDER-EM` like algorithm, and these results will be specialized in section 9. For all $i = 1, \ldots, n$, $\bar{s}_i \circ \mathsf{T}$ is a function from $\mathbb{R}^q$ to $\mathbb{R}^q$; for a selection of b indices $\mathcal{B}$ in $\{1, \ldots, n\}$ with or without replacement, we set $\bar{s}_{\mathcal{B}} \circ \mathsf{T} \stackrel{\text{def}}{=} \mathsf{b}^{-1} \sum_{i \in \mathcal{B}} \bar{s}_i \circ \mathsf{T}$. More generally, $\bar{s} \circ \mathsf{T} \stackrel{\text{def}}{=} n^{-1} \sum_{i=1}^n \bar{s}_i \circ \mathsf{T}$. For some results below, specific assumptions may be introduced on $\bar{s}_t \circ \mathsf{T}$.

Let $\{\gamma_k, k \geq 1\}$ be a positive deterministic sequence. Let $\{\mathcal{B}_k, k \geq 1\}$ be a family of independent random mini batches sampled in $\{1, \ldots, n\}$ of size b, (either with replacement or without replacement). Finally, let U_{-1}, U_0 be random variables. Assume that (U_{-1}, U_0) are independent from the sequences $\{\mathcal{B}_k, k \geq 1\}$ and set

$$\widetilde{U}_0 \stackrel{\text{def}}{=} \bar{s} \circ \mathsf{T}(U_{-1}) = \mathbb{E}\left[\bar{s}_{\mathcal{B}_1} \circ \mathsf{T}(U_{-1}) | U_{-1} \right] \; . \tag{16}$$

Consider the recursive definition for $k \geq 0$,

$$\begin{aligned} \widetilde{U}_{k+1} &= \widetilde{U}_k + \bar{s}_{\mathcal{B}_{k+1}} \circ \mathsf{T}(U_k) - \bar{s}_{\mathcal{B}_{k+1}} \circ \mathsf{T}(U_{k-1}) \; , \\ U_{k+1} &= U_k + \gamma_{k+1} \left(\widetilde{U}_{k+1} - U_k \right) \; . \end{aligned}$$

Finally, define the filtration

$$\mathcal{G}_0 \stackrel{\text{def}}{=} \sigma(U_{-1}, U_0), \qquad \text{for } k \geq 0, \; \mathcal{G}_{k+1} \stackrel{\text{def}}{=} \sigma\left(\mathcal{G}_k \cup \mathcal{B}_{k+1} \right) \; ,$$

and define the sequence of random variables

$$\Delta_0 \stackrel{\text{def}}{=} h(U_{-1}), \qquad \text{for } k \geq 0, \; \Delta_{k+1} \stackrel{\text{def}}{=} \widetilde{U}_{k+1} - U_k = \gamma_{k+1}^{-1} (U_{k+1} - U_k) \; .$$

Lemma 4. *For any $k \geq 0$, $\mathcal{B}_{k+1}$ and $\mathcal{G}_k$ are independent. For any $u \in \mathbb{R}^q$,*

$$\mathbb{E}\left[\bar{s}_{\mathcal{B}_{k+1}} \circ \mathsf{T}(u) \right] = \bar{s} \circ \mathsf{T}(u) \; .$$

Assume that $\bar{s}_i \circ \mathsf{T}$ is globally Lipschitz with constant L_i; set $L^2 \stackrel{\text{def}}{=} n^{-1} \sum_{i=1}^n L_i^2$. For any $u, u' \in \mathbb{R}^q$,

$$\begin{aligned} &\mathbb{E}\left[\| \bar{s}_{\mathcal{B}_{k+1}} \circ \mathsf{T}(u) - \bar{s}_{\mathcal{B}_{k+1}} \circ \mathsf{T}(u') - \bar{s} \circ \mathsf{T}(u) + \bar{s} \circ \mathsf{T}(u') \|^2 \right] \\ &\qquad \leq \frac{1}{\mathsf{b}} \left(L^2 \|u - u'\|^2 - \| \bar{s} \circ \mathsf{T}(u) - \bar{s} \circ \mathsf{T}(u') \|^2 \right) \; . \end{aligned}$$

Proof. By assumption, $\mathcal{B}_{k+1}$ and (U_0, U_{-1}) are independent, and therefore $\mathcal{B}_{k+1}$ and $\mathcal{G}_0$ are also. In addition, $\mathcal{B}_{k+1}$ is independent of $\mathcal{B}_\ell$ for any $\ell \leq k$ so $\mathcal{B}_{k+1}$ is independent of $\mathcal{G}_k$.

• Case: sampling with replacement. We write $\mathcal{B}_{k+1} = \{I_1, \cdots, I_{\mathsf{b}}\}$ where the random variables are independent, and uniformly distributed on $\{1, \cdots, n\}$. Then

$$\mathbb{E}\left[\bar{s}_{\mathcal{B}_{k+1}} \circ \mathsf{T}(u)\right] = \frac{1}{\mathsf{b}} \sum_{\ell=1}^{\mathsf{b}} \mathbb{E}\left[\bar{s}_{I_\ell} \circ \mathsf{T}(u)\right] = \mathbb{E}\left[\bar{s}_{I_1} \circ \mathsf{T}(u)\right] = \bar{s} \circ \mathsf{T}(u) \; .$$

In addition, since the variance of the sum is the sum of the variance for independent r.v.

$$\begin{aligned}
&\mathbb{E}\left[\|\bar{s}_{\mathcal{B}_{k+1}} \circ \mathsf{T}(u) - \bar{s}_{\mathcal{B}_{k+1}} \circ \mathsf{T}(u') - \bar{s} \circ \mathsf{T}(u) + \bar{s} \circ \mathsf{T}(u')\|^2\right] \\
&\quad = \frac{1}{\mathsf{b}^2} \sum_{\ell=1}^{\mathsf{b}} \mathbb{E}\left[\|\bar{s}_{I_\ell} \circ \mathsf{T}(u) - \bar{s}_{I_\ell} \circ \mathsf{T}(u') - \bar{s} \circ \mathsf{T}(u) + \bar{s} \circ \mathsf{T}(u')\|^2\right]
\end{aligned}$$

Then we have

$$\begin{aligned}
&\mathbb{E}\left[\|\bar{s}_{I_\ell} \circ \mathsf{T}(u) - \bar{s}_{I_\ell} \circ \mathsf{T}(u') - \bar{s} \circ \mathsf{T}(u) + \bar{s} \circ \mathsf{T}(u')\|^2\right] \\
&\quad = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[\|\bar{s}_i \circ \mathsf{T}(u) - \bar{s}_i \circ \mathsf{T}(u')\|^2\right] - \|\bar{s} \circ \mathsf{T}(u) + \bar{s} \circ \mathsf{T}(u')\|^2 \\
&\quad \leq \|u - u'\|^2 \frac{1}{n} \sum_{i=1}^{n} L_i^2 - \|\bar{s} \circ \mathsf{T}(u) + \bar{s} \circ \mathsf{T}(u')\|^2 \qquad (17)
\end{aligned}$$

which concludes the proof.

• Case: sampling with no replacement. I_1 is a uniform random variable on $\{1, \cdots, n\}$ so that $\mathbb{E}\left[\bar{s}_{I_1} \circ \mathsf{T}(u)\right] = \bar{s} \circ \mathsf{T}(u)$. Conditionally to I_1, I_2 is a uniform random variable on $\{1, \cdots, n\} \backslash \{I_1\}$. Therefore

$$\mathbb{E}\left[\bar{s}_{I_2} \circ \mathsf{T}(u)\right] = \frac{1}{n-1} \left(\sum_{j=1}^{n} \bar{s}_j \circ \mathsf{T}(u) - \mathbb{E}\left[\bar{s}_{I_1} \circ \mathsf{T}(u)\right] \right) = \frac{n}{n-1} \bar{s} \circ \mathsf{T}(u) - \frac{1}{n-1} \bar{s} \circ \mathsf{T}(u) \; .$$

By induction, for any $\ell \geq 2$,

$$\begin{aligned}
\mathbb{E}\left[\bar{s}_{I_\ell} \circ \mathsf{T}(u)\right] &= \frac{1}{n-\ell+1} \left(\sum_{j=1}^{n} \bar{s}_j \circ \mathsf{T}(u) - \sum_{q=1}^{\ell-1} \mathbb{E}\left[\bar{s}_{I_q} \circ \mathsf{T}(u)\right] \right) \\
&= \frac{n}{n-\ell+1} \bar{s} \circ \mathsf{T}(u) - \frac{\ell-1}{n-\ell+1} \bar{s} \circ \mathsf{T}(u) \; .
\end{aligned}$$

As a conclusion, $\mathsf{b}^{-1} \sum_{\ell=1}^{\mathsf{b}} \mathbb{E}\left[\bar{s}_{I_\ell} \circ \mathsf{T}(u)\right] = \bar{s} \circ \mathsf{T}(u)$. Let $u, u' \in \mathbb{R}^q$; set $\phi(I_\ell) \stackrel{\text{def}}{=} \bar{s}_{I_\ell} \circ \mathsf{T}(u) - \bar{s} \circ \mathsf{T}(u) + \bar{s}_{I_\ell} \circ \mathsf{T}(u') - \bar{s} \circ \mathsf{T}(u')$. Then $\mathbb{E}\left[\phi(I_\ell)\right] = 0$. We first prove by induction that $\mathbb{E}\left[\|\phi(I_\ell)\|^2\right] = \mathbb{E}\left[\|\phi(I_1)\|^2\right]$. Upon noting that I_1 is a uniform random variable on $\{1, \cdots, n\}$,

$$\begin{aligned}
\mathbb{E}\left[\|\phi(I_\ell)\|^2\right] &= \frac{1}{n-\ell+1} \left(\sum_{i=1}^{n} \|\phi(i)\|^2 - \mathbb{E}\left[\|\phi(I_1)\|^2 + \cdots + \|\phi(I_{\ell-1})\|^2\right] \right) \\
&= \frac{n}{n-\ell+1} \mathbb{E}\left[\|\phi(I_1)\|^2\right] - \frac{1}{n-\ell+1} \sum_{p=1}^{\ell-1} \mathbb{E}\left[\|\phi(I_p)\|^2\right]
\end{aligned}$$

which concludes the induction. Second, let us prove that for any $\ell \geq 0$,

$$\mathbb{E}\left[\| \sum_{p=1}^{\ell+1} \phi(I_p) \|^2 \right] \leq (\ell+1) \mathbb{E}\left[\|\phi(I_1)\|^2\right] \; . \qquad (18)$$

Since $n^{-1} \sum_{i=1}^{n} \phi(i) = \mathbb{E}\left[\phi(I_1)\right] = 0$,

$$\mathbb{E}\left[\left\langle \sum_{p=1}^{\ell} \phi(I_p), \phi(I_{\ell+1}) \right\rangle \right] = \frac{1}{n-\ell} \mathbb{E}\left[\left\langle \sum_{p=1}^{\ell} \phi(I_p), \sum_{i=1}^{n} \phi(i) - \sum_{p=1}^{\ell} \phi(I_p) \right\rangle \right] = -\frac{1}{n-\ell} \mathbb{E}\left[\| \sum_{p=1}^{\ell} \phi(I_p) \|^2 \right] \; ,$$

so that

$$\mathbb{E}\left[\|\sum_{p=1}^{\ell+1}\phi(I_p)\|^2\right] = \left(1-\frac{2}{n-\ell}\right)\mathbb{E}\left[\|\sum_{p=1}^{\ell}\phi(I_p)\|^2\right] + \mathbb{E}\left[\|\phi(I_{\ell+1})\|^2\right] \le (\ell+1)\mathbb{E}\left[\|\phi(I_1)\|^2\right] .$$

The proof follows from (18) and (17) since here again, I_1 is uniformly distributed on $\{1,\cdots,n\}$. □

Lemma 5. *For any $k \ge 0$,*

$$\mathbb{E}\left[\Delta_{k+1}|\mathcal{G}_k\right] - h(U_k) = \Delta_k - h(U_{k-1}) .$$

Proof. Let $k \ge 0$. Since conditionally to $\mathcal{G}_k$, $\mathcal{B}_{k+1} = \{I_1,\ldots,I_{\mathsf{b}}\}$ where the random variables I_k's are independent and uniformly distributed on $\{1,\ldots,n\}$, we have

$$\mathbb{E}\left[\widetilde{U}_{k+1}|\mathcal{G}_k\right] = \widetilde{U}_k + \bar{s}\circ\mathsf{T}(U_k) - \bar{s}\circ\mathsf{T}(U_{k-1}) .$$

In the case $k=0$, we have by using (16)

$$\mathbb{E}\left[\Delta_1 - h(U_0)|\mathcal{G}_0\right] = \mathbb{E}\left[\widetilde{U}_1|\mathcal{G}_0\right] - \bar{s}\circ\mathsf{T}(U_0) = 0 = \Delta_0 - h(U_{-1}) ;$$

the last equality explains the convention for Δ_0. In the case $k>0$,

$$\begin{aligned}\mathbb{E}\left[\Delta_{k+1}|\mathcal{G}_k\right] &= \mathbb{E}\left[\widetilde{U}_{k+1} - U_k|\mathcal{G}_k\right] = \widetilde{U}_k + h(U_k) - \bar{s}\circ\mathsf{T}(U_{k-1}) \\ &= \Delta_k + U_{k-1} + h(U_k) - \bar{\circ}\mathsf{T}(U_{k-1}) = h(U_k) + \Delta_k - h(U_{k-1}) .\end{aligned}$$

□

Proposition 6. *Assume that for all $i=1,\cdots,n$, $\bar{s}_i\circ\mathsf{T}$ is globally Lipschitz, with constant L_i; set $L^2 \stackrel{\text{def}}{=} n^{-1}\sum_{i=1}^n L_i^2$. Then $\Delta_0 - \mathbb{E}\left[\Delta_0|\mathcal{G}_0\right] = 0$,*

$$\begin{aligned}\mathbb{E}[\|\Delta_1 - \mathbb{E}\left[\Delta_1|\mathcal{G}_0\right]\|^2|\mathcal{G}_0] &= \mathbb{E}[\|\Delta_1 - h(U_0)\|^2|\mathcal{G}_0] \\ &\le -\frac{1}{\mathsf{b}}\|\bar{s}\circ\mathsf{T}(U_0) - \bar{s}\circ\mathsf{T}(U_{-1})\|^2 + \frac{L^2}{\mathsf{b}}\|U_0 - U_{-1}\|^2 .\end{aligned}$$

and for any $k \ge 1$,

$$\mathbb{E}[\|\Delta_{k+1} - \mathbb{E}[\Delta_{k+1}|\mathcal{G}_k]\|^2|\mathcal{G}_k] \le -\frac{1}{\mathsf{b}}\|\bar{s}\circ\mathsf{T}(U_k) - \bar{s}\circ\mathsf{T}(U_{k-1})\|^2 + \frac{L^2}{\mathsf{b}}\gamma_k^2\|\Delta_k\|^2 ;$$

$$\begin{aligned}\mathbb{E}[\|\Delta_{k+1} - h(U_k)\|^2|\mathcal{G}_0] &\le -\frac{1}{\mathsf{b}}\sum_{j=0}^{k}\mathbb{E}\left[\|\bar{s}\circ\mathsf{T}(U_j) - \bar{s}\circ\mathsf{T}(U_{j-1})\|^2|\mathcal{G}_0\right] \\ &\quad + \frac{L^2}{\mathsf{b}}\left(\sum_{j=1}^{k}\gamma_j^2\,\mathbb{E}\left[\|\Delta_j\|^2|\mathcal{G}_0\right] + \|U_0 - U_{-1}\|^2\right) .\end{aligned}$$

Proof. The statement on Δ_0 is trivial since $\Delta_0 = h(U_{-1}) \in \mathcal{G}_0$. By definition of Δ_1, by Lemma 4 and by (16)

$$\mathbb{E}\left[\Delta_1|\mathcal{G}_0\right] = \mathbb{E}\left[\widetilde{U}_1|\mathcal{G}_0\right] - U_0 = \widetilde{U}_0 + \bar{s}\circ\mathsf{T}(U_0) - \bar{s}\circ\mathsf{T}(U_{-1}) - U_0 = h(U_0) .$$

The equation

$$\Delta_1 - \mathbb{E}\left[\Delta_1|\mathcal{G}_0\right] = \bar{s}_{\mathcal{B}_1}\circ\mathsf{T}(U_0) - \bar{s}_{\mathcal{B}_1}\circ\mathsf{T}(U_{-1}) - (\bar{s}\circ\mathsf{T}(U_0) - \bar{s}\circ\mathsf{T}(U_{-1}))$$

and Lemma 4 provides the upper bound for Δ_1. Let $k \ge 1$. By definition of Δ_{k+1} and by Lemma 4,

$$\begin{aligned}\Delta_{k+1} - \mathbb{E}\left[\Delta_{k+1}|\mathcal{G}_k\right] &= \widetilde{U}_{k+1} - \mathbb{E}\left[\widetilde{U}_{k+1}|\mathcal{G}_k\right] \\ &= \bar{s}_{\mathcal{B}_{k+1}}\circ\mathsf{T}(U_k) - \bar{s}_{\mathcal{B}_{k+1}}\circ\mathsf{T}(U_{k-1}) + \bar{s}\circ\mathsf{T}(U_k) - \bar{s}\circ\mathsf{T}(U_{k-1})\end{aligned}$$

and we then conclude by Lemma 4 again. For the second statement, since we have $\mathbb{E}\left[\|U\|^2\right] = \mathbb{E}\left[\|U - \mathbb{E}[U|V]\|^2\right] + \mathbb{E}\left[\|\mathbb{E}[U|V]\|^2\right]$ for any random variables U, V, it holds for any $k \geq 0$,

$$\begin{aligned}\mathbb{E}\left[\|\Delta_{k+1} - h(U_k)\|^2|\mathcal{G}_k\right] &= \mathbb{E}\left[\|\Delta_{k+1} - \mathbb{E}\left[\Delta_{k+1}|\mathcal{G}_k\right]\|^2|\mathcal{G}_k\right] + \|\mathbb{E}\left[\Delta_{k+1}|\mathcal{G}_k\right] - h(U_k)\|^2 \\ &= \mathbb{E}\left[\|\Delta_{k+1} - \mathbb{E}\left[\Delta_{k+1}|\mathcal{G}_k\right]\|^2|\mathcal{G}_k\right] + \|\Delta_k - h(U_{k-1})\|^2\end{aligned}$$

where we used Lemma 5 in the last equality. By induction, this yields

$$\mathbb{E}\left[\|\Delta_{k+1} - h(U_k)\|^2|\mathcal{G}_0\right] = \sum_{j=0}^{k} \mathbb{E}\left[\mathbb{E}\left[\|\Delta_{j+1} - \mathbb{E}\left[\Delta_{j+1}|\mathcal{G}_j\right]\|^2|\mathcal{G}_j\right]|\mathcal{G}_0\right]$$

where we have used that $\Delta_0 - h(U_{-1}) = 0$ (by definition). We then conclude with the first statement. □

Lemma 7. *For any $h, s, S \in \mathbb{R}^q$ and any $q \times q$ symmetric matrix B, it holds*

$$-2\langle Bh, S\rangle = -\langle BS, S\rangle - \langle Bh, h\rangle + \langle B\{h - S\}, h - S\rangle \ .$$

Proposition 8. *Assume H1, H2, H3 and H4 and H5. It holds for any $K \geq 2$,*

$$\begin{aligned}&\sum_{\ell=1}^{K-1} \delta_\ell \, \mathbb{E}\left[\|U_\ell - U_{\ell-1}\|^2|\mathcal{G}_0\right] + \frac{v_{\min}}{2} \sum_{k=0}^{K-2} \gamma_{k+1} \mathbb{E}\left[\|h(U_k)\|^2|\mathcal{G}_0\right] \\ &\qquad \leq \mathrm{W}(U_0) - \mathbb{E}\left[\mathrm{W}(U_{K-1})|\mathcal{G}_0\right] + \frac{L^2 v_{\max}}{2\mathsf{b}} \left(\sum_{k=1}^{K-1} \gamma_k\right) \|U_0 - U_{-1}\|^2 \ ,\end{aligned}$$

where (by convention, $\sum_{\ell=K-1}^{K-2} = 0$)

$$\delta_\ell \stackrel{\text{def}}{=} \left(\frac{v_{\min}}{2\gamma_\ell} - \frac{L_{\nabla \mathrm{W}}}{2} - \frac{v_{\max}}{2} \frac{L^2}{\mathsf{b}} \sum_{k=\ell}^{K-2} \gamma_{k+1}\right)$$

Proof. Let $k \in \{0, \cdots, K-2\}$. By Proposition 1 and H5-Item (c), W is continuously differentiable with globally Lipschitz gradient, which implies

$$\mathrm{W}(U_{k+1}) - \mathrm{W}(U_k) \leq \langle \nabla \mathrm{W}(U_k), U_{k+1} - U_k\rangle + \frac{L_{\nabla \mathrm{W}}}{2}\|U_{k+1} - U_k\|^2 \ .$$

By Proposition 1, we have $\nabla \mathrm{W}(U_k) = -B(U_k)h(U_k)$; hence,

$$\langle \nabla \mathrm{W}(U_k), U_{k+1} - U_k\rangle = -\langle B(U_k)h(U_k), U_{k+1} - U_k\rangle \ .$$

We apply Lemma 7 with $B \leftarrow B(U_k)$, $h \leftarrow h(U_k)$ and $S \leftarrow \Delta_{k+1} = (U_{k+1} - U_k)/\gamma_{k+1}$. This yields by H5-Item (a),

$$\langle \nabla \mathrm{W}(U_k), U_{k+1} - U_k\rangle \leq -\frac{\gamma_{k+1} v_{\min}}{2}\|\Delta_{k+1}\|^2 - \frac{v_{\min}\gamma_{k+1}}{2}\|h(U_k)\|^2 + \frac{v_{\max}\gamma_{k+1}}{2}\|h(U_k) - \Delta_{k+1}\|^2$$

and since $\Delta_{k+1} = (U_{k+1} - U_k)/\gamma_{k+1}$, we obtain

$$\langle \nabla \mathrm{W}(U_k), U_{k+1} - U_k\rangle \leq -\frac{v_{\min}}{2\gamma_{k+1}}\|U_{k+1} - U_k\|^2 - \frac{v_{\min}\gamma_{k+1}}{2}\|h(U_k)\|^2 + \frac{v_{\max}\gamma_{k+1}}{2}\|\Delta_{k+1} - h(U_k)\|^2 \ .$$

Therefore, we established

$$\begin{aligned}\left(\frac{v_{\min}}{2\gamma_{k+1}} - \frac{L_{\nabla \mathrm{W}}}{2}\right)\|U_{k+1} - U_k\|^2 + \frac{v_{\min}\gamma_{k+1}}{2}\|h(U_k)\|^2 &\leq \frac{v_{\max}\gamma_{k+1}}{2}\|\Delta_{k+1} - h(U_k)\|^2 \\ &\quad + \mathrm{W}(U_k) - \mathrm{W}(U_{k+1}) \ .\end{aligned}$$

Applying the conditional expectation and using Proposition 6 (and again $\gamma_j^2\|\Delta_j\|^2 = \|U_j - U_{j-1}\|^2$ for $j \geq 1$), this yields

$$\begin{aligned}&\left(\frac{v_{\min}}{2\gamma_{k+1}} - \frac{L_{\nabla \mathrm{W}}}{2}\right)\mathbb{E}\left[\|U_{k+1} - U_k\|^2|\mathcal{G}_0\right] + \frac{v_{\min}\gamma_{k+1}}{2}\mathbb{E}\left[\|h(U_k)\|^2|\mathcal{G}_0\right] \\ &\qquad \leq \frac{v_{\max}\gamma_{k+1}}{2}\frac{L^2}{\mathsf{b}}\sum_{j=0}^{k}\mathbb{E}\left[\|U_j - U_{j-1}\|^2|\mathcal{G}_0\right] + \mathbb{E}\left[\mathrm{W}(U_k) - \mathrm{W}(U_{k+1})|\mathcal{G}_0\right] \ .\end{aligned}$$

We now sum from $k = 0$ to $k = K-2$ and obtain by using Lemma 9 with $\bar{\Delta}_j \leftarrow \mathbb{E}\left[\|U_j - U_{j-1}\|^2 | \mathcal{G}_0\right]$,

$$\begin{aligned}
&\left(\frac{v_{\min}}{2\gamma_{K-1}} - \frac{L_{\nabla \mathrm{W}}}{2}\right) \mathbb{E}\left[\|U_{K-1} - U_{K-2}\|^2 | \mathcal{G}_0\right] \\
&+ \sum_{\ell=1}^{K-2} \left(\frac{v_{\min}}{2\gamma_\ell} - \frac{L_{\nabla \mathrm{W}}}{2} - \frac{v_{\max}}{2}\frac{L^2}{\mathsf{b}} \sum_{k=\ell}^{K-2} \gamma_{k+1}\right) \mathbb{E}\left[\|U_\ell - U_{\ell-1}\|^2 | \mathcal{G}_0\right] \\
&+ \frac{v_{\min}}{2} \sum_{k=0}^{K-2} \gamma_{k+1} \mathbb{E}\left[\|h(U_k)\|^2 | \mathcal{G}_0\right] \le \mathbb{E}\left[\mathrm{W}(U_0) - \mathrm{W}(U_{K-1}) | \mathcal{G}_0\right] \\
&+ \|U_0 - U_{-1}\|^2 \left(\sum_{k=1}^{K-1} \gamma_k\right) \frac{L^2 v_{\max}}{2\mathsf{b}} \, .
\end{aligned}$$

This concludes the proof. □

Lemma 9. *For any real numbers $a_i, b_i, \bar{\Delta}_i$ and $K \ge 2$,*

$$\sum_{k=1}^{K-1} \left(a_k \bar{\Delta}_k - b_k \sum_{\ell=0}^{k-1} \bar{\Delta}_\ell\right) = a_{K-1}\bar{\Delta}_{K-1} - \bar{\Delta}_0 \sum_{k=1}^{K-1} b_k + \sum_{\ell=1}^{K-2} \left(a_\ell - \sum_{k=\ell+1}^{K-1} b_k\right) \bar{\Delta}_\ell \, .$$

Lemma 10. *For any $k \ge (t-1)k_{\mathrm{in}}$,*

$$\begin{aligned}
&\sum_{q=(t-1)k_{\mathrm{in}}}^{k} \left(-a_{q+1}X_{q+1} + b_{q+1} \sum_{j=(t-1)k_{\mathrm{in}}}^{q} Y_j + c_{q+1} \sum_{j=(t-1)k_{\mathrm{in}}}^{q} d_j X_j\right) \\
&\qquad = -a_{k+1}X_{k+1} + d_{(t-1)k_{\mathrm{in}}} \left(\sum_{q=(t-1)k_{\mathrm{in}}}^{k} c_{q+1}\right) X_{(t-1)k_{\mathrm{in}}} \\
&\qquad + \sum_{j=(t-1)k_{\mathrm{in}}+1}^{k} \left(d_j \left(\sum_{q=j}^{k} c_{q+1}\right) - a_j\right) X_j + \sum_{j=(t-1)k_{\mathrm{in}}}^{k} \left(\sum_{q=j}^{k} b_{q+1}\right) Y_j \, .
\end{aligned}$$

9 Proof of Main Results in section 3

For $t = 1, \cdots, k_{\mathrm{out}}$ and $k = 0, \cdots, k_{\mathrm{in}} - 2$, define the σ-field $\mathcal{F}_{t,k}$:

$$\mathcal{F}_{0,k_{\mathrm{in}}-1} \stackrel{\mathrm{def}}{=} \sigma(\widehat{S}_{\mathrm{init}}) \, , \qquad \mathcal{F}_{t,0} \stackrel{\mathrm{def}}{=} \mathcal{F}_{t-1,k_{\mathrm{in}}-1} \, , \qquad \mathcal{F}_{t,k+1} \stackrel{\mathrm{def}}{=} \sigma\left(\mathcal{F}_{t,k} \cup \mathcal{B}_{t,k+1}\right) \, .$$

With these definitions, we have for $t = 1, \cdots, k_{\mathrm{out}}$ and $k = 0, \cdots, k_{\mathrm{in}} - 2$,

$$\widehat{S}_{t,k+1} \in \mathcal{F}_{t,k+1} \, , \qquad \mathsf{S}_{t,k+1} \in \mathcal{F}_{t,k+1} \, , \qquad \mathcal{B}_{t,k+1} \in \mathcal{F}_{t,k+1} \, ;$$

and $\widehat{S}_{t,0} \in \mathcal{F}_{t,0}$, $\mathsf{S}_{t,0} \in \mathcal{F}_{t,0}$. For $t = 1, \cdots, k_{\mathrm{out}}$ and $k = 0, \cdots, k_{\mathrm{in}} - 2$ set

$$H_{t,k+1} \stackrel{\mathrm{def}}{=} \gamma_{t,k+1}^{-1} \left(\widehat{S}_{t,k+1} - \widehat{S}_{t,k}\right) = \mathsf{S}_{t,k+1} - \widehat{S}_{t,k} \in \mathcal{F}_{t,k+1} \, ; \tag{19}$$

and choose the convention $H_{1,0} \stackrel{\mathrm{def}}{=} h(\widehat{S}_{1,-1})$, and

$$H_{t+1,0} = H_{t,k_{\mathrm{in}}} \stackrel{\mathrm{def}}{=} \gamma_{t,k_{\mathrm{in}}}^{-1} (\widehat{S}_{t+1,0} - \widehat{S}_{t,k_{\mathrm{in}}-1}) = \mathsf{S}_{t+1,0} - \widehat{S}_{t,k_{\mathrm{in}}-1} = h\left(\widehat{S}_{t,k_{\mathrm{in}}-1}\right) \, . \tag{20}$$

9.1 Preliminary lemmas

The following results are consequences of the general analysis in section 8.

Lemma 11. *Assume H1, H2, H3. Let* $\{\widehat{S}_{t,k}, t = 1, \cdots, k_{\rm out}, k = 0, \cdots, k_{\rm in} - 1\}$ *be the sequence given by algorithm 1. For* $t = 1, \cdots, k_{\rm out}$ *and* $k = 0, \cdots, k_{\rm in} - 2$

$$\mathbb{E}\left[H_{t,k+1}|\mathcal{F}_{t,k}\right] - h(\widehat{S}_{t,k}) = H_{t,k} - h(\widehat{S}_{t,k-1}) \; ,$$
$$H_{t,0} - h(\widehat{S}_{t,-1}) = 0 = H_{t,k_{\rm in}} - h(\widehat{S}_{t,k_{\rm in}-1}) \; .$$

Proof. Let $t \geq 1$: apply Lemma 5 with $U_0 \leftarrow \widehat{S}_{t,0}$, $U_{-1} \leftarrow \widehat{S}_{t,-1}$, $\gamma_{k+1} \leftarrow \gamma_{t,k+1}$, $\mathcal{B}_{k+1} \leftarrow \mathcal{B}_{t,k+1}$. Then $\widetilde{U}_0 \leftarrow \mathsf{S}_{t,0}$ satisfies the condition (16) and for any $k \geq 0$, we have $U_{k+1} = \widehat{S}_{t,k+1}$, $\widetilde{U}_{k+1} = \mathsf{S}_{t,k+1}$, $\Delta_{k+1} = H_{t,k+1}$ and $\mathcal{G}_{k+1} = \mathcal{F}_{t,k+1}$. This yields the result. □

Corollary 12 (of Lemma 11)**.** *For* $t = 1, \cdots, k_{\rm out}$ *and* $k = 0, \cdots, k_{\rm in}$

$$\mathbb{E}[H_{t,k} - h(\widehat{S}_{t,k-1})|\mathcal{F}_{t,0}] = 0 \; .$$

Proof. Let $t \geq 1$. If $k = 0$ then by Lemma 11, the property holds. Let $k \in \{0, \ldots, k_{\rm in} - 2\}$. We write by using Lemma 11

$$\mathbb{E}[H_{t,k+1} - h(\widehat{S}_{t,k})|\mathcal{F}_{t,0}] = \mathbb{E}[\mathbb{E}[H_{t,k+1} - h(\widehat{S}_{t,k})|\mathcal{F}_{t,k}]|\mathcal{F}_{t,0}] = \mathbb{E}[H_{t,k} - h(\widehat{S}_{t,k-1})|\mathcal{F}_{t,0}] \; .$$

The proof is concluded by induction:

$$\mathbb{E}[H_{t,k+1} - h(\widehat{S}_{t,k})|\mathcal{F}_{t,0}] = \mathbb{E}[H_{t,0} - h(\widehat{S}_{t,-1})|\mathcal{F}_{t,0}] = 0 \; .$$

□

Proposition 13. *Assume H1, H2, H3, H5-(b) and set* $L^2 \stackrel{\rm def}{=} n^{-1}\sum_{i=1}^n L_i^2$. *For any* $t = 1, \cdots, k_{\rm out}$, $H_{t,0} - h(\widehat{S}_{t,-1}) = 0$, *and*

$$\mathbb{E}[\|H_{t,1} - \mathbb{E}\left[H_{t,1}|\mathcal{F}_{t,0}\right]\|^2|\mathcal{F}_{t,0}] \leq -\frac{1}{\mathsf{b}}\|\bar{s} \circ \mathsf{T}(\widehat{S}_{t,0}) - \bar{s} \circ \mathsf{T}(\widehat{S}_{t,-1})\|^2 + \frac{L^2}{\mathsf{b}}\|\widehat{S}_{t,0} - \widehat{S}_{t,-1}\|^2 \; .$$

In addition, for $k = 1, \cdots, k_{\rm in} - 2$,

$$\mathbb{E}[\|H_{t,k+1} - h(\widehat{S}_{t,k})\|^2|\mathcal{F}_{t,0}] \leq -\frac{1}{\mathsf{b}}\sum_{j=0}^{k}\mathbb{E}\left[\|\bar{s} \circ \mathsf{T}(\widehat{S}_{t,j}) - \bar{s} \circ \mathsf{T}(\widehat{S}_{t,j-1})\|^2|\mathcal{F}_{t,0}\right]$$
$$+ \frac{L^2}{\mathsf{b}}\left(\sum_{j=1}^{k}\gamma_{t,j}^2\,\mathbb{E}\left[\|H_{t,j}\|^2|\mathcal{F}_{t,0}\right] + \|\widehat{S}_{t,0} - \widehat{S}_{t,-1}\|^2\right) \; ,$$

$$\mathbb{E}[\|H_{t,k+1} - \mathbb{E}[H_{t,k+1}|\mathcal{F}_{t,k}]\|^2|\mathcal{F}_{t,k}] \leq -\frac{1}{\mathsf{b}}\|\bar{s} \circ \mathsf{T}(\widehat{S}_{t,k}) - \bar{s} \circ \mathsf{T}(\widehat{S}_{t,k-1})\|^2 + \frac{L^2}{\mathsf{b}}\gamma_{t,k}^2\,\|H_{t,k}\|^2 \; .$$

Finally,

$$\|H_{t,k_{\rm in}} - h(\widehat{S}_{t,k_{\rm in}-1})\| = \|H_{t,k_{\rm in}} - \mathbb{E}\left[H_{t,k_{\rm in}}|\mathcal{F}_{t,k_{\rm in}-1}\right]\| = 0 \; .$$

Proof. Let $t \geq 1$. Apply Proposition 6 with $\gamma_k \leftarrow \gamma_{t,k}$, $\mathcal{B}_{k+1} \leftarrow \mathcal{B}_{t,k+1}$, $U_0 \leftarrow \widehat{S}_{t,0}$, $U_{-1} \leftarrow \widehat{S}_{t,-1}$, $\mathcal{G}_k \leftarrow \mathcal{F}_{t,k}$. Since $\mathsf{S}_{t,0} = \bar{s} \circ \mathsf{T}(\widehat{S}_{t,-1})$, then the condition (16) is satisfied with $\widetilde{U}_0 = \mathsf{S}_{t,0}$. Conclude by observing that $\widetilde{U}_k = \mathsf{S}_{t,k}$ and $\Delta_{k+1} = H_{t,k+1}$. □

9.2 Proof of Theorem 2

Proposition 14. *Assume H1, H2, H3, H4 and H5. Set* $L^2 \stackrel{\rm def}{=} n^{-1}\sum_{i=1}^n L_i^2$. *For any positive numbers* $\beta_{t,k}$, *set for* $t = 1, \cdots, k_{\rm out}$ *and* $k = 0, \cdots, k_{\rm in} - 1$

$$A_{t,k} \stackrel{\rm def}{=} \gamma_{t,k} v_{\min}\left(1 - \frac{\beta_{t,k}^2}{2v_{\min}} - \gamma_{t,k}\frac{L_{\nabla \rm W}}{2v_{\min}} - \frac{L^2 v_{\max}^2}{2v_{\min}\mathsf{b}}\gamma_{t,k}\left(\sum_{\ell=k}^{k_{\rm in}-2}\frac{\gamma_{t,\ell+1}}{\beta_{t,\ell+1}^2}\right)\right)$$
$$B_{t,k} \stackrel{\rm def}{=} \frac{v_{\max}^2}{2\mathsf{b}}\sum_{k=0}^{k_{\rm in}-2}\left(\sum_{\ell=k}^{k_{\rm in}-2}\frac{\gamma_{t,\ell+1}}{\beta_{t,\ell+1}^2}\right) \; ;$$

by convention $\beta_{t,0}=0$, $\gamma_{t,0}=\gamma_{t-1,k_{\rm in}}$, $\gamma_{0,k_{\rm in}}=0$ *and* $B_{t,k_{\rm in}-1}=0$.

Let $\{\widehat{S}_{t,k}, t=1,\cdots,k_{\rm out}; k=0,\cdots,k_{\rm in}-1\}$ *be the sequence given by algorithm 1. For any* $t=1,\cdots,k_{\rm out}$,

$$\mathrm{W}(\widehat{S}_{t,0}) \le \mathrm{W}(\widehat{S}_{t,-1}) - \gamma_{t-1,k_{\rm in}} v_{\min}\left(1-\gamma_{t-1,k_{\rm in}}\frac{L_{\nabla\mathrm{W}}}{2v_{\min}}\right)\|h(\widehat{S}_{t,-1})\|^2 \; ; \tag{21}$$

and

$$\sum_{t=1}^{k_{\rm out}}\sum_{k=0}^{k_{\rm in}-1}\left(A_{t,k}\mathbb{E}[\|H_{t,k}\|^2] + B_{t,k}\mathbb{E}[\|\bar{s}\circ\mathsf{T}(\widehat{S}_{t,k})-\bar{s}\circ\mathsf{T}(\widehat{S}_{t,k-1})\|^2]\right) \le \mathbb{E}[\mathrm{W}(\widehat{S}_{\rm init})] - \min \mathrm{W} \; .$$

Proof. Let $t\ge 1$. By H5-(c), we have for any $k=-1,\cdots,k_{\rm in}-1$,

$$\mathrm{W}(\widehat{S}_{t,k+1}) \le \mathrm{W}(\widehat{S}_{t,k}) + \gamma_{t,k+1}\left\langle \nabla\mathrm{W}(\widehat{S}_{t,k}), H_{t,k+1}\right\rangle + \gamma_{t,k+1}^2\frac{L_{\nabla\mathrm{W}}}{2}\|H_{t,k+1}\|^2 \; ; \tag{22}$$

by convention, we set $\widehat{S}_{t,k_{\rm in}} \stackrel{\rm def}{=} \widehat{S}_{t+1,0}$. By Proposition 1, H5-(a) and (20), we have

$$\left\langle \nabla\mathrm{W}(\widehat{S}_{t,k_{\rm in}-1}), H_{t,k_{\rm in}}\right\rangle \le -v_{\min}\|h(\widehat{S}_{t,k_{\rm in}-1})\|^2 = -v_{\min}\|H_{t,k_{\rm in}}\|^2 \; ,$$

so that

$$\mathrm{W}(\widehat{S}_{t,k_{\rm in}}) \le \mathrm{W}(\widehat{S}_{t,k_{\rm in}-1}) - \gamma_{t,k_{\rm in}} v_{\min}\|H_{t,k_{\rm in}}\|^2 + \gamma_{t,k_{\rm in}}^2\frac{L_{\nabla\mathrm{W}}}{2}\|H_{t,k_{\rm in}}\|^2 \; . \tag{23}$$

This concludes the proof of (21) since $\widehat{S}_{t,k_{\rm in}} = \widehat{S}_{t+1,0}$ and $\widehat{S}_{t,k_{\rm in}-1} = \widehat{S}_{k+1,-1}$. Now, let us fix $k\in\{0,\cdots,k_{\rm in}-2\}$. We write

$$\begin{aligned}
\left\langle \nabla\mathrm{W}(\widehat{S}_{t,k}), H_{t,k+1}\right\rangle &= -\left\langle B(\widehat{S}_{t,k})h(\widehat{S}_{t,k}), H_{t,k+1}\right\rangle \\
&= -\left\langle B(\widehat{S}_{t,k})\left(h(\widehat{S}_{t,k})-H_{t,k+1}\right), H_{t,k+1}\right\rangle - \left\langle B(\widehat{S}_{t,k})H_{t,k+1}, H_{t,k+1}\right\rangle \\
&\le -\left\langle B(\widehat{S}_{t,k})\left(h(\widehat{S}_{t,k})-H_{t,k+1}\right), H_{t,k+1}\right\rangle - v_{\min}\|H_{t,k+1}\|^2 \; .
\end{aligned} \tag{24}$$

Note that for $a,b\in\mathbb{R}^q$ and $\beta>0$,

$$\langle a,b\rangle \le \frac{\beta^2}{2}\|a\|^2 + \frac{1}{2\beta^2}\|b\|^2 \; .$$

By H5-(a), we have for any $\beta_{t,k+1}>0$,

$$\left|\left\langle B(\widehat{S}_{t,k})\left(h(\widehat{S}_{t,k})-H_{t,k+1}\right), H_{t,k+1}\right\rangle\right| \le \frac{\beta_{t,k+1}^2}{2}\|H_{t,k+1}\|^2 + \frac{v_{\max}^2}{2\beta_{t,k+1}^2}\|H_{t,k+1}-h(\widehat{S}_{t,k})\|^2 \; . \tag{25}$$

Combining (22), (24) and (25) yield

$$\mathrm{W}(\widehat{S}_{t,k+1}) \le \mathrm{W}(\widehat{S}_{t,k}) - \Lambda_{t,k+1}\|H_{t,k+1}\|^2 + \gamma_{t,k+1}\frac{v_{\max}^2}{2\beta_{t,k+1}^2}\|H_{t,k+1}-h(\widehat{S}_{t,k})\|^2 \; ,$$

where for $\ell=1,\ldots,k_{\rm in}-1$,

$$\Lambda_{t,\ell} \stackrel{\rm def}{=} \gamma_{t,\ell}v_{\min}\left(1-\frac{\beta_{t,\ell}^2}{2v_{\min}} - \gamma_{t,\ell}\frac{L_{\nabla\mathrm{W}}}{2v_{\min}}\right) \; .$$

By Proposition 13,

$$\begin{aligned}
\mathbb{E}\left[\mathrm{W}(\widehat{S}_{t,k+1})|\mathcal{F}_{t,0}\right] \le{}& \mathbb{E}\left[\mathrm{W}(\widehat{S}_{t,k})|\mathcal{F}_{t,0}\right] - \Lambda_{t,k+1}\mathbb{E}\left[\|H_{t,k+1}\|^2|\mathcal{F}_{t,0}\right] \\
&- \gamma_{t,k+1}\frac{v_{\max}^2}{2\beta_{t,k+1}^2}\frac{1}{\mathsf{b}}\sum_{j=0}^{k}\mathbb{E}\left[\|\bar{s}\circ\mathsf{T}(\widehat{S}_{t,j})-\bar{s}\circ\mathsf{T}(\widehat{S}_{t,j-1})\|^2|\mathcal{F}_{t,0}\right] \\
&+ \gamma_{t,k+1}\frac{v_{\max}^2}{2\beta_{t,k+1}^2}\frac{L^2}{\mathsf{b}}\left(\sum_{j=1}^{k}\gamma_{t,j}^2\,\mathbb{E}\left[\|H_{t,j}\|^2|\mathcal{F}_{t,0}\right] + \|\widehat{S}_{t,0}-\widehat{S}_{t,-1}\|^2\right) \; ;
\end{aligned}$$

by taking the expectation, this yields

$$\begin{aligned}\mathbb{E}\left[\mathrm{W}(\widehat{S}_{t,k+1})\right] &\leq \mathbb{E}\left[\mathrm{W}(\widehat{S}_{t,k})\right] - \Lambda_{t,k+1}\mathbb{E}\left[\|H_{t,k+1}\|^2\right] \\ &\quad - \gamma_{t,k+1}\frac{v_{\max}^2}{2\beta_{t,k+1}^2}\frac{1}{\mathsf{b}}\sum_{j=0}^{k}\mathbb{E}\left[\|\bar{s}\circ\mathsf{T}(\widehat{S}_{t,j}) - \bar{s}\circ\mathsf{T}(\widehat{S}_{t,j-1})\|^2\right] \\ &\quad + \gamma_{t,k+1}\frac{v_{\max}^2}{2\beta_{t,k+1}^2}\frac{L^2}{\mathsf{b}}\left(\sum_{j=1}^{k}\gamma_{t,j}^2\,\mathbb{E}\left[\|H_{t,j}\|^2\right] + \mathbb{E}\left[\|\widehat{S}_{t,0} - \widehat{S}_{t,-1}\|^2\right]\right)\ ;\end{aligned}$$

By summing from time $k=0$ to $k=k_{\mathrm{in}}-2$, we have (see Lemma 10)

$$\begin{aligned}\mathbb{E}\left[\mathrm{W}(\widehat{S}_{t+1,-1})\right] = \mathbb{E}\left[\mathrm{W}(\widehat{S}_{t,k_{\mathrm{in}}-1})\right] &\leq \mathbb{E}\left[\mathrm{W}(\widehat{S}_{t,0})\right] - \Lambda_{t,k_{\mathrm{in}}-1}\mathbb{E}\left[\|H_{t,k_{\mathrm{in}}-1}\|^2\right] \\ &\quad + \frac{v_{\max}^2L^2}{2\mathsf{b}}\left(\sum_{\ell=0}^{k_{\mathrm{in}}-2}\frac{\gamma_{t,\ell+1}}{\beta_{t,\ell+1}^2}\right)\mathbb{E}\left[\|\widehat{S}_{t,0} - \widehat{S}_{t,-1}\|^2\right] \\ &\quad - \frac{v_{\max}^2}{2\mathsf{b}}\sum_{k=0}^{k_{\mathrm{in}}-2}\left(\sum_{\ell=k}^{k_{\mathrm{in}}-2}\frac{\gamma_{t,\ell+1}}{\beta_{t,\ell+1}^2}\right)\mathbb{E}\left[\|\bar{s}\circ\mathsf{T}(\widehat{S}_{t,k}) - \bar{s}\circ\mathsf{T}(\widehat{S}_{t,k-1})\|^2\right] \\ &\quad + \sum_{k=1}^{k_{\mathrm{in}}-2}\left(\frac{L^2v_{\max}^2}{2\mathsf{b}}\gamma_{t,k}^2\left(\sum_{\ell=k}^{k_{\mathrm{in}}-2}\frac{\gamma_{t,\ell+1}}{\beta_{t,\ell+1}^2}\right) - \Lambda_{t,k}\right)\mathbb{E}\left[\|H_{t,k}\|^2\right]\ .\end{aligned}$$

With (21), and using $H_{t,k_{\mathrm{in}}} = h(\widehat{S}_{t,k_{\mathrm{in}}-1}) = h(\widehat{S}_{t+1,-1})$; $\widehat{S}_{1,0} = \widehat{S}_{1,-1} = \widehat{S}_{\mathrm{init}}$; and for $t\geq 2$, $\widehat{S}_{t,0} - \widehat{S}_{t,-1} = \gamma_{t-1,k_{\mathrm{in}}}h(\widehat{S}_{t-1,k_{\mathrm{in}}-1}) = \gamma_{t-1,k_{\mathrm{in}}}H_{t-1,k_{\mathrm{in}}} = \gamma_{t-1,k_{\mathrm{in}}}H_{t,0}$:

$$\begin{aligned}&\mathbb{E}\left[\mathrm{W}(\widehat{S}_{t+1,0})\right] - \mathbb{E}\left[\mathrm{W}(\widehat{S}_{t,0})\right] \\ &\leq -\Lambda_{t,k_{\mathrm{in}}-1}\mathbb{E}\left[\|H_{t,k_{\mathrm{in}}-1}\|^2\right] + \frac{v_{\max}^2L^2}{2\mathsf{b}}\gamma_{t-1,k_{\mathrm{in}}}^2\left(\sum_{\ell=0}^{k_{\mathrm{in}}-2}\frac{\gamma_{t,\ell+1}}{\beta_{t,\ell+1}^2}\right)\mathbb{E}\left[\|H_{t,0}\|^2\right]1_{t>1} \\ &- \frac{v_{\max}^2}{2\mathsf{b}}\sum_{k=0}^{k_{\mathrm{in}}-2}\left(\sum_{\ell=k}^{k_{\mathrm{in}}-2}\frac{\gamma_{t,\ell+1}}{\beta_{t,\ell+1}^2}\right)\mathbb{E}\left[\|\bar{s}\circ\mathsf{T}(\widehat{S}_{t,k}) - \bar{s}\circ\mathsf{T}(\widehat{S}_{t,k-1})\|^2\right] \\ &+ \sum_{k=1}^{k_{\mathrm{in}}-2}\left(\frac{L^2v_{\max}^2}{2\mathsf{b}}\gamma_{t,k}^2\left(\sum_{\ell=k}^{k_{\mathrm{in}}-2}\frac{\gamma_{t,\ell+1}}{\beta_{t,\ell+1}^2}\right) - \Lambda_{t,k}\right)\mathbb{E}\left[\|H_{t,k}\|^2\right] - \gamma_{t,k_{\mathrm{in}}}v_{\min}\left(1-\gamma_{t,k_{\mathrm{in}}}\frac{L_{\nabla\mathrm{W}}}{2v_{\min}}\right)\mathbb{E}\left[\|H_{t+1,0}\|^2\right] \\ &\leq -B_{t,k}\mathbb{E}\left[\|\bar{s}\circ\mathsf{T}(\widehat{S}_{t,k}) - \bar{s}\circ\mathsf{T}(\widehat{S}_{t,k-1})\|^2\right] + \sum_{k=1}^{k_{\mathrm{in}}-1}\left(\frac{L^2v_{\max}^2}{2\mathsf{b}}\gamma_{t,k}^2\left(\sum_{\ell=k}^{k_{\mathrm{in}}-2}\frac{\gamma_{t,\ell+1}}{\beta_{t,\ell+1}^2}\right) - \Lambda_{t,k}\right)\mathbb{E}\left[\|H_{t,k}\|^2\right] \\ &+ \frac{v_{\max}^2L^2}{2\mathsf{b}}\gamma_{t-1,k_{\mathrm{in}}}^2\left(\sum_{\ell=0}^{k_{\mathrm{in}}-2}\frac{\gamma_{t,\ell+1}}{\beta_{t,\ell+1}^2}\right)\mathbb{E}\left[\|H_{t,0}\|^2\right]1_{t>1} - \gamma_{t,k_{\mathrm{in}}}v_{\min}\left(1-\gamma_{t,k_{\mathrm{in}}}\frac{L_{\nabla\mathrm{W}}}{2v_{\min}}\right)\mathbb{E}\left[\|H_{t+1,0}\|^2\right]\ .\end{aligned}$$

We now sum from $t=1$ to $t=k_{\mathrm{out}}$. □

Corollary 15 (of Proposition 14)**.** *Choose $\alpha>0$, $\beta>0$ such that*

$$C(\alpha,\beta) \stackrel{\mathrm{def}}{=} 1 - \frac{\beta^2}{2v_{\min}} - \frac{\alpha}{2v_{\min}}\frac{L_{\nabla\mathrm{W}}}{L} - \frac{\alpha^2v_{\max}^2}{2\beta^2v_{\min}}\frac{k_{\mathrm{in}}}{\mathsf{b}}$$

is positive; and set

$$\gamma_{t,k+1} \stackrel{\mathrm{def}}{=} \frac{\alpha}{L}\ , \qquad \beta_{t,k+1} \stackrel{\mathrm{def}}{=} \beta\ .$$

Then, for uniform random variables τ,ξ on $\{1,\cdots,k_{\mathrm{out}}\}$ and $\{0,\cdots,k_{\mathrm{in}}-1\}$ respectively, independent from $\mathcal{F}_{k_{\mathrm{out}},k_{\mathrm{in}}-1}$,

$$\mathbb{E}\left[\|H_{\tau,\xi}\|^2\right] \leq \frac{L}{\alpha v_{\min}C(\alpha,\beta)}\frac{1}{k_{\mathrm{in}}k_{\mathrm{out}}}\left(\mathbb{E}\left[\mathrm{W}(\widehat{S}_{\mathrm{init}})\right] - \min\mathrm{W}\right)\ .$$

Proof. We have

$$A_{t,k} \geq \frac{\alpha v_{\min}}{L}\left(1 - \frac{\beta^2}{2v_{\min}} - \frac{\alpha}{2v_{\min}}\frac{L_{\nabla \mathrm{W}}}{L} - \frac{\alpha^2 v_{\max}^2}{2\beta^2 v_{\min}}\frac{k_{\mathrm{in}}}{\mathsf{b}}\right) ,$$
$$B_{t,k} \geq \frac{v_{\max}^2}{2\mathsf{b}}\frac{\alpha}{L\beta^2}k_{\mathrm{in}} ,$$

from which the conclusion follows. □

Proof of Theorem 2 Let τ, ξ be uniform random variables resp. on $\{1, \cdots, k_{\mathrm{out}}\}$ and $\{0, \cdots, k_{\mathrm{in}} - 1\}$. Since $\widehat{S}_{1,-1} = \widehat{S}_{1,0}$ and for $t \geq 2$, $\widehat{S}_{t,-1} = \widehat{S}_{t-1,k_{\mathrm{in}}-1}$, then $\widehat{S}_{t,\xi-1}$ is well defined. We write

$$\mathbb{E}\left[\|h(\widehat{S}_{\tau,\xi-1})\|^2\right] \leq 2\mathbb{E}\left[\|H_{\tau,\xi}\|^2\right] + 2\,\mathbb{E}\left[\|H_{\tau,\xi} - h(\widehat{S}_{\tau,\xi-1})\|^2\right] .$$

For the second term, we have

$$\mathbb{E}[\|H_{\tau,\xi} - h(\widehat{S}_{\tau,\xi-1})\|^2] = \frac{1}{k_{\mathrm{in}}k_{\mathrm{out}}}\sum_{t=1}^{k_{\mathrm{out}}}\sum_{k=0}^{k_{\mathrm{in}}-1}\mathbb{E}[\|H_{t,k} - h(\widehat{S}_{t,k-1})\|^2] \tag{26}$$

by Proposition 13, since $\widehat{S}_{1,0} = \widehat{S}_{1,-1}$, the RHS of (26) is upper bounded by

$$\frac{\alpha^2}{\mathsf{b}}\frac{1}{k_{\mathrm{out}}}\sum_{t=1}^{k_{\mathrm{out}}}\sum_{k=0}^{k_{\mathrm{in}}-1}\mathbb{E}\left[\|H_{t,k}\|^2\right] \leq \frac{\alpha^2 k_{\mathrm{in}}}{\mathsf{b}}\mathbb{E}\left[\|H_{\tau,\xi}\|^2\right] .$$

The proof is concluded by Corollary 15:

$$\mathbb{E}[\|h(\widehat{S}_{\tau,\xi-1})\|^2] \leq \left(\frac{1}{k_{\mathrm{in}}} + \frac{\alpha^2}{\mathsf{b}}\right)\frac{2L}{\alpha v_{\min}C(\alpha,\beta)}\frac{1}{k_{\mathrm{out}}}\left(\mathbb{E}[\mathrm{W}(\widehat{S}_{\mathrm{init}})] - \min \mathrm{W}\right) . \tag{27}$$

Let us choose $\beta > 0$ so that $\beta \mapsto C(\alpha, \beta)$ is maximal: for $A, B > 0$, the function $x \mapsto x/A + B/x$ is minimal at $x_\star \stackrel{\mathrm{def}}{=} \sqrt{AB}$. This yields

$$\beta^2(\alpha) \stackrel{\mathrm{def}}{=} \alpha v_{\max}\sqrt{\frac{k_{\mathrm{in}}}{\mathsf{b}}} ,$$

and

$$v_{\min}\, C(\alpha, \beta(\alpha)) \stackrel{\mathrm{def}}{=} v_{\min} - \alpha\mu_\star , \qquad \mu_\star \stackrel{\mathrm{def}}{=} v_{\max}\sqrt{\frac{k_{\mathrm{in}}}{\mathsf{b}}} + \frac{L_{\nabla \mathrm{W}}}{2L} .$$

The function $\alpha \mapsto \alpha\, v_{\min}\, C(\alpha, \beta(\alpha))$ is maximal when $\alpha_\star \stackrel{\mathrm{def}}{=} v_{\min}/(2\mu_\star)$ thus yielding $\alpha_\star\, v_{\min}\, C(\alpha_\star, \beta(\alpha_\star)) = v_{\min}^2/(4\mu_\star)$. By replacing $\beta \leftarrow \beta(\alpha)$ and $\alpha \leftarrow \alpha_\star$ in (27), we have

$$\mathbb{E}[\|h(\widehat{S}_{\tau,\xi-1})\|^2] \leq \left(\mu_\star + \frac{k_{\mathrm{in}}v_{\min}^2}{4\mu_\star \mathsf{b}}\right)\frac{8L}{v_{\min}^2}\frac{1}{k_{\mathrm{in}}\,k_{\mathrm{out}}}\left(\mathbb{E}[\mathrm{W}(\widehat{S}_{\mathrm{init}})] - \min \mathrm{W}\right) . \tag{28}$$

9.3 On the Batch Size b and Epoch Length k_{in}

Assume that $\mathsf{b} = O(n^{\mathsf{a}})$ and $k_{\mathrm{in}} = O(n^{\mathsf{c}})$ for some $\mathsf{a}, \mathsf{c} \geq 0$. Let $\epsilon > 0$.

Case $\mathsf{a} \geq \mathsf{c}$. When $n \to \infty$, $\mu_\star(k_{\mathrm{in}}, \mathsf{b}) = O(1)$. Choose $\alpha \in (0, v_{\min}/\mu_\star(k_{\mathrm{in}}, \mathsf{b}))$ such that $\alpha = O(n^{-\mathsf{d}})$ for some $\mathsf{d} \geq 0$.

The RHS in (15) is lower than ϵ by choosing

$$k_{\mathrm{out}} = O\left(\epsilon^{-1}n^{-\mathsf{c}}\left(n^{\mathsf{d}} + \frac{1}{n^{\mathsf{d}+\mathsf{a}-\mathsf{c}}}\right)\right) ;$$

this implies that

$$K_{\mathrm{CE}}(n, \epsilon) = O\left(n + (n + n^{\mathsf{a}+\mathsf{c}})k_{\mathrm{out}}\right) , \qquad K_{\mathrm{Opt}}(n, \epsilon) = O\left(1 + (1 + n^{\mathsf{c}})k_{\mathrm{out}}\right) .$$

In order to make k_{out} as small as possible, we choose $\mathsf{d} = 0$ and c as large as possible (i.e. $\mathsf{a} = \mathsf{c}$). Hence $k_{\text{out}} = O(\epsilon^{-1} n^{-\mathsf{a}})$. This implies that $K_{\text{Opt}}(n, \epsilon) = O(\epsilon^{-1})$. For fixed $\mathsf{a} \geq 0$, $K_{\text{CE}}(n, \epsilon)$ is optimized by choosing $\mathsf{a} \leq 1 - \mathsf{a}$, which implies $\mathsf{a} \leq 1/2$. The largest value of a will provide the best rate for k_{out}. Hence, the conclusion is

$$\mathsf{a} = \mathsf{c} = 1/2, \qquad \mathsf{d} = 0,$$

which yields $\mathsf{b} = O(\sqrt{n})$, $k_{\text{in}} = O(\sqrt{n})$, $k_{\text{out}} = O(\epsilon^{-1} n^{-1/2})$, $K_{\text{CE}}(n, \epsilon) = O(n + \epsilon^{-1}\sqrt{n})$ and $K_{\text{Opt}}(n, \epsilon) = O(\epsilon^{-1})$.

Case $\mathsf{a} < \mathsf{c}$. When $n \to \infty$, $\mu_\star(k_{\text{in}}, \mathsf{b}) = O(n^{(\mathsf{c}-\mathsf{a})/2})$. Choose $\alpha \in (0, v_{\min}/\mu_\star(k_{\text{in}}, \mathsf{b}))$ such that $\alpha = O(n^{-\mathsf{d}})$ for some $\mathsf{d} \geq (\mathsf{c} - \mathsf{a})/2$.

The RHS in (15) is lower than ϵ by choosing

$$k_{\text{out}} = O\left(\epsilon^{-1} n^{-\mathsf{c}} \left(n^{\mathsf{d}} + \frac{1}{n^{\mathsf{d}+\mathsf{a}-\mathsf{c}}}\right)\right) ;$$

we also have

$$K_{\text{CE}}(n, \epsilon) = O\left(n + (n + n^{\mathsf{a}+\mathsf{c}}) k_{\text{out}}\right) , \qquad K_{\text{Opt}}(n, \epsilon) = O\left(1 + (1 + n^{\mathsf{c}}) k_{\text{out}}\right) .$$

In order to make k_{out} as small as possible, we choose $\mathsf{d} = (\mathsf{c} - \mathsf{a})/2$ so $k_{\text{out}} = O(\epsilon^{-1} n^{-(\mathsf{a}+\mathsf{c})/2})$, and then we choose $\mathsf{c}+\mathsf{a}$ as large as possible. Hence This implies that $K_{\text{Opt}}(n, \epsilon) = O(\epsilon^{-1} n^{(\mathsf{c}-\mathsf{a})/2})$ and $K_{\text{Opt}}(n, \epsilon)$ is optimized by choosing $\mathsf{c} - \mathsf{a}$ as small as possible. Finally, $K_{\text{CE}}(n, \epsilon)$ is optimized with $\mathsf{a} + \mathsf{c} \leq 1$. Hence, the conclusion is: choose $\delta > 0$ and set

$$\mathsf{a} = (1 - \delta)/2, \qquad \mathsf{c} = (1 + \delta)/2, \qquad \mathsf{d} = \delta/2,$$

which yields $\mathsf{b} = O(n^{1/2-\delta/2})$, $k_{\text{in}} = O(n^{1/2+\delta/2})$, $k_{\text{out}} = O(\epsilon^{-1} n^{-1/2})$, $K_{\text{CE}}(n, \epsilon) = O(n + \epsilon^{-1}\sqrt{n})$ and $K_{\text{Opt}}(n, \epsilon) = O(\epsilon^{-1} n^{\delta/2})$.

Conclusion. The above discussion shows that the best complexity in terms of the number of computations of per-sample conditional expectations and the one in terms of number of parameter updates are both optimized in the case $\mathsf{a} = \mathsf{c} = 1/2$.

10 Linear convergence rate of `SPIDER-EM-PL`

In this section, we establish a linear convergence rate of a slightly modified version of `SPIDER-EM`, see algorithm 8, the main modification being in the initialization. The proof is adapted from [27, Theorem 5].

Data: $k_{\text{in}} \in \mathbb{N}_\star$, $k_{\text{out}} \in \mathbb{N}_\star$, $\widehat{S}_{\text{init}} \in \mathbb{R}^q$, $\{\gamma_{t,k+1}, t = 1, \cdots, k_{\text{out}} \text{ and } k = 0, \cdots, k_{\text{in}} - 1\}$ positive sequence.

Result: A `SPIDER-EM-PL` sequence: $\widehat{S}_{t,k}, t = 1, \cdots, k_{\text{out}}, k = 0, \ldots, k_{\text{in}} - 1$

$\mathsf{S}_{1,0} = \bar{s} \circ \mathsf{T}(\widehat{S}_{\text{init}})$, $\widehat{S}_{1,0} = \widehat{S}_{1,-1} = \widehat{S}_{\text{init}}$;

for $t = 1, \ldots, k_{\text{out}}$ **do**

  Sample ξ_t a uniform random variable on $\{1, \cdots, k_{\text{in}} - 1\}$;

  for $k = 0, \cdots, \xi_t - 1$ **do**

    Sample a mini-batch $\mathcal{B}_{t,k+1}$ in $\{1, \ldots, n\}$ of size b, with or without replacement ;

    $\mathsf{S}_{t,k+1} = \mathsf{S}_{t,k} + \bar{s}_{\mathcal{B}_{t,k+1}} \circ \mathsf{T}(\widehat{S}_{t,k}) - \bar{s}_{\mathcal{B}_{t,k+1}} \circ \mathsf{T}(\widehat{S}_{t,k-1})$;

    $\widehat{S}_{t,k+1} = \widehat{S}_{t,k} + \gamma_{t,k+1}\big(\mathsf{S}_{t,k+1} - \widehat{S}_{t,k}\big)$

  $\widehat{S}_{t+1,0} = \widehat{S}_{t+1,-1} = \widehat{S}_{t,\xi_t}$;

  $\mathsf{S}_{t+1,0} = \bar{s} \circ \mathsf{T}(\widehat{S}_{t,\xi_t})$

Algorithm 8: The `SPIDER-EM-PL` algorithm.

By Proposition 8, we have

Proposition 16. *Assume H1, H2, H3 and H4 and H5. Set* $L^2 \overset{\text{def}}{=} n^{-1}\sum_{i=1}^n L_i^2$. *For any integers* $t \geq 1$ *and* $K \geq 2$

$$\sum_{\ell=1}^{K-1} \delta_{t,\ell}\, \mathbb{E}\left[\|\widehat{S}_{t,\ell} - \widehat{S}_{t,\ell-1}\|^2 | \mathcal{F}_{t,0}\right] + \frac{v_{\min}}{2} \sum_{k=0}^{K-2} \gamma_{t,k+1} \mathbb{E}\left[\|h(\widehat{S}_{t,k})\|^2 | \mathcal{F}_{t,0}\right] \leq \mathbb{E}\left[\mathrm{W}(\widehat{S}_{t,0}) - \mathrm{W}(\widehat{S}_{t,K-1}) | \mathcal{F}_{t,0}\right] ,$$

where (by convention, $\sum_{\ell=K-1}^{K-2} = 0$*),*

$$\delta_{t,\ell} \overset{\text{def}}{=} \left(\frac{v_{\min}}{2\gamma_{t,\ell}} - \frac{L_{\nabla \mathrm{W}}}{2} - \frac{v_{\max}}{2} \frac{L^2}{\mathsf{b}} \sum_{k=\ell}^{K-2} \gamma_{t,k+1} \right) .$$

Corollary 17 (of Proposition 16)**.** *For any* $\gamma > 0$ *such that*

$$\gamma^2 + \frac{L_{\nabla \mathrm{W}}\, \mathsf{b}}{v_{\max} L^2 (K-1)} \gamma - \frac{v_{\min} \mathsf{b}}{v_{\max} L^2 (K-1)} < 0 ,$$

we have

$$\frac{v_{\min}\gamma}{2} \sum_{k=0}^{K-1} \mathbb{E}\left[\|h(\widehat{S}_{t,k})\|^2 | \mathcal{F}_{t,0}\right] \leq \mathbb{E}\left[\mathrm{W}(\widehat{S}_{t,0}) - \mathrm{W}(\widehat{S}_{t,K}) | \mathcal{F}_{t,0}\right] .$$

As a consequence of Corollary 17, if ξ_t is a uniform random variable on $\{1, \cdots, k_{\mathrm{in}}-1\}$ independent of the other random variables, then

$$\mathbb{E}\left[\|h(\widehat{S}_{t,\xi_t})\|^2\right] \leq \frac{2}{v_{\min}\gamma(k_{\mathrm{in}}-1)} \mathbb{E}\left[\mathrm{W}(\widehat{S}_{t,0}) - \min \mathrm{W}\right] .$$

When the Polyak-Lojasiewicz inequality holds

$$\exists \tau^\star > 0 \text{ such that } \forall s, \mathrm{W}(s) - \min \mathrm{W} \leq \tau^\star \|\nabla \mathrm{W}(s)\|^2 , \tag{29}$$

this yields by H5-Item (a)

$$\mathbb{E}\left[\|h(\widehat{S}_{t,\xi_t})\|^2\right] \leq \frac{2}{v_{\min}\gamma(k_{\mathrm{in}}-1)} \mathbb{E}\left[\mathrm{W}(\widehat{S}_{t,0}) - \min \mathrm{W}\right] \leq \frac{2\tau^\star v_{\max}^2}{v_{\min}\gamma(k_{\mathrm{in}}-1)} \mathbb{E}\left[\|h(\widehat{S}_{t,0})\|^2\right] .$$

The above discussion establishes the following result.

Theorem 18. *Assume H1, H2, H3, H4 and H5 and set* $L^2 \overset{\text{def}}{=} n^{-1}\sum_{i=1}^n L_i^2$. *Assume also that the Polyak-Lojasiewicz inequality (29) holds. Fix* $k_{\mathrm{out}}, k_{\mathrm{in}} \in \mathbb{N}_\star$, $\mathsf{b} \in \mathbb{N}_\star$*; set* $\gamma_{t,k+1} \overset{\text{def}}{=} \gamma$ *for any* $t \geq 1, k \geq 0$ *for some* $\gamma > 0$ *satisfying*

$$\gamma^2 + \frac{L_{\nabla \mathrm{W}}\, \mathsf{b}}{v_{\max} L^2 (k_{\mathrm{in}}-1)} \gamma - \frac{v_{\min} \mathsf{b}}{v_{\max} L^2 (k_{\mathrm{in}}-1)} < 0 .$$

Let $\{\widehat{S}_{t,k}, t = 1, \cdots, k_{\mathrm{out}}, k = 0, \cdots, \xi_t\}$ *be the sequence given by algorithm 8. Then*

$$\mathbb{E}\left[\|h(\widehat{S}_{t+1,0})\|^2\right] = \mathbb{E}\left[\|h(\widehat{S}_{t,\xi_t})\|^2\right] \leq \frac{2\tau^\star v_{\max}^2}{v_{\min}\gamma(k_{\mathrm{in}}-1)} \mathbb{E}\left[\|h(\widehat{S}_{t,0})\|^2\right] .$$

11 Mixture of Gaussian distributions

In this section, we use the common notation $\{\widehat{S}_\ell, \ell \geq 0\}$ for a path. For `sEM-vr` and `SPIDER-EM`, $\widehat{S}_\ell$ stands for $\widehat{S}_{t_\ell,k_\ell}$ where $t_\ell \geq 1$ and $k_\ell \in \{0, \cdots, k_{\mathrm{in}}-1\}$ are the unique integers such that $\ell = (t_\ell - 1)k_{\mathrm{in}} + k_\ell$.

11.1 The model

Consider a mixture of Gaussian distributions on $\mathbb{R}^p$,

$$y \mapsto \sum_{\ell=1}^{g} \alpha_\ell \, \mathcal{N}_p\left(\mu_\ell, \Sigma\right)[y] \; ; \tag{30}$$

$\mathcal{N}_p\left(\mu_\ell, \Sigma\right)[y]$ denotes the density of a $\mathbb{R}^p$-valued Gaussian distribution with expectation μ_ℓ, covariance matrix Σ and evaluated at $y \in \mathbb{R}^p$. We consider a parametric statistical model indexed by $\theta \stackrel{\text{def}}{=} (\alpha_1, \ldots, \alpha_g, \mu_1, \ldots, \mu_g, \Sigma)$ in Θ where

$$\Theta \stackrel{\text{def}}{=} \left\{ \alpha_\ell \geq 0, \sum_{\ell=1}^{g} \alpha_\ell = 1 \right\} \times \mathbb{R}^{pg} \times \mathcal{M}_p^+ \; ; \tag{31}$$

$\mathcal{M}_p^+$ denotes the set of positive definite $p \times p$ matrices.

Given n examples $y_1, \ldots, y_n$ modeled as independent realizations of a mixture of Gaussian distributions as described by (30), the log-likelihood is

$$\theta \mapsto \sum_{i=1}^{n} \log \sum_{\ell=1}^{g} \alpha_\ell \, \mathcal{N}_p\left(\mu_\ell, \Sigma\right)[y_i] \; .$$

Proposition 19 shows that the minimization of the negative log-likelihood on Θ is covered by the optimization problem addressed in the paper.

Proposition 19. *Set $\Gamma \stackrel{\text{def}}{=} \Sigma^{-1}$, and define for $y \in \mathbb{R}^p$ and $z \in \{1, \ldots, g\}$,*

$$\mathsf{A}_y \stackrel{\text{def}}{=} \begin{bmatrix} \mathrm{I}_g \\ \mathrm{I}_g \otimes y \end{bmatrix} \in \mathbb{R}^{g(1+p) \times g} \; , \qquad \rho(z) \stackrel{\text{def}}{=} \begin{bmatrix} 1_{z=1} \\ \cdots \\ 1_{z=g} \end{bmatrix} \; .$$

The negative normalized log-likelihood is of the form (2) with $\rho(y,z) = 1$, $s(y,z) \stackrel{\text{def}}{=} \mathsf{A}_y \, \rho(z)$ and

$$\phi(\theta) \stackrel{\text{def}}{=} \begin{bmatrix} \ln \alpha_1 - 0.5 \mu_1^T \Gamma \mu_1 \\ \cdots \\ \ln \alpha_g - 0.5 \mu_g^T \Gamma \mu_g \\ \Gamma \mu_1 \\ \cdots \\ \Gamma \mu_g \end{bmatrix} \; , \tag{32}$$

$$\psi(\theta) \stackrel{\text{def}}{=} \frac{p}{2} \ln(2\pi) + \frac{1}{2} \mathrm{Tr}\left(\frac{\Gamma}{n} \sum_{i=1}^{n} y_i y_i^T \right) - \frac{1}{2} \ln \det(\Gamma) \; . \tag{33}$$

Proof. The likelihood of a single observation y_i is given by

$$\begin{aligned}
\theta &\mapsto \frac{1}{\sqrt{2\pi}^p} \sum_{z=1}^{g} \alpha_z \sqrt{\det(\Gamma)} \exp\left(-\frac{1}{2} (y_i - \mu_z)^T \Gamma (y_i - \mu_z) \right) \\
&= \frac{\sqrt{\det(\Gamma)}}{\sqrt{2\pi}^p} \exp\left(-\frac{1}{2} y_i^T \Gamma y_i \right) \sum_{z=1}^{g} \exp\left(\sum_{\ell=1}^{g} 1_{z=\ell} \left\{ \ln \alpha_\ell - 0.5 \mu_\ell^T \Gamma \mu_\ell + \mu_\ell^T \Gamma y_i \right\} \right) \\
&= \frac{\sqrt{\det(\Gamma)}}{\sqrt{2\pi}^p} \exp\left(-\frac{1}{2} \mathrm{Tr}(\Gamma y_i y_i^T) \right) \sum_{z=1}^{g} \exp\left(\sum_{\ell=1}^{g} 1_{z=\ell} \{ \ln \alpha_\ell - 0.5 \mu_\ell^T \Gamma \mu_\ell \} + \sum_{\ell=1}^{g} \langle \Gamma \mu_\ell, y_i 1_{z=\ell} \rangle \right) \\
&= \frac{\sqrt{\det(\Gamma)}}{\sqrt{2\pi}^p} \exp\left(-\frac{1}{2} \mathrm{Tr}(\Gamma y_i y_i^T) \right) \sum_{z=1}^{g} \exp\left(\langle s(y_i, z), \phi(\theta) \rangle \right)
\end{aligned}$$

where we used that $\mathrm{Tr}(Auu^T) = u^T A u$. Since the observations are modeled as independent, the log-likelihood of the n observations $y_1, \ldots, y_n$ is

$$\theta \mapsto \frac{n}{2} \left(\log \det(\Gamma) - p \log(2\pi) \right) - \frac{1}{2} \mathrm{Tr}(\Gamma \sum_{i=1}^{n} y_i y_i^T) + \sum_{i=1}^{n} \log \sum_{z=1}^{g} \exp\left(\langle s(y_i, z), \phi(\theta) \rangle \right) \; .$$

This yields the expression of the negative normalized log-likeliood. ∎

The following statement gives the expression of the optimization map T. It relies on standard computations; the proof is omitted.

Proposition 20. *Let ϕ, ψ and Θ resp. given by Proposition 19 and (31). For any $s = (s_1, \ldots, s_{g+pg}) \in \mathbb{R}^{g+pg}$ in the set*

$$\left(s_1 > 0, \ldots, s_g > 0, \frac{1}{n} \sum_{i=1}^n y_i y_i^T - \sum_{\ell=1}^g s_\ell^{-1} s_{g+(\ell-1)p+1:g+\ell p} \, s_{g+(\ell-1)p+1:g+\ell p}^T \text{ positive definite} \right)$$

the minimizer of $\theta \mapsto -\langle s, \phi(\theta) \rangle + \psi(\theta)$ under the constraint that $\theta \in \Theta$, exists and is unique and is given by

$$\begin{aligned}
\alpha_\ell &\stackrel{\text{def}}{=} \frac{s_\ell}{\sum_{u=1}^g s_u} \, , \qquad \ell = 1, \ldots, g \, , \\
\mu_\ell &\stackrel{\text{def}}{=} \frac{1}{s_\ell} s_{g+(\ell-1)p+1:g+\ell p} \, , \qquad \ell = 1, \ldots, g \, , \\
\Sigma^{-1} &\stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^n y_i y_i^T - \sum_{\ell=1}^g s_\ell \mu_\ell \mu_\ell^T \, .
\end{aligned}$$

Proposition 21 provides the expression of the conditional probabilities $z \mapsto p(z|y_i;\theta)$ on $\{1, \ldots, g\}$; as a corollary of this statement, we also have the expression of the per sample conditional expectations

$$\bar{s}_i(\theta) \stackrel{\text{def}}{=} \sum_{z=1}^g s(y_i, z) \, p(z|y_i;\theta) \, ,$$

for all $i = 1, \ldots, n$.

Proposition 21. *For any $y \in \mathbb{R}^p$, $z \in \{1, \ldots, g\}$ and $\theta \in \Theta$ where Θ is defined by (31), we have*

$$p(z|y;\theta) \stackrel{\text{def}}{=} \frac{\alpha_z \, \mathcal{N}_p(\mu_z, \Sigma)[y]}{\sum_{u=1}^g \alpha_u \, \mathcal{N}_p(\mu_u, \Sigma)[y]} \, , \tag{34}$$

and

$$\sum_{z=1}^g s(y, z) \, p(z|y;\theta) = \begin{bmatrix} p(1|y;\theta) \\ \ldots \\ p(g|y;\theta) \\ y \, p(1|y;\theta) \\ \ldots \\ y \, p(g|y;\theta) \end{bmatrix} \, ,$$

where $s(y, z)$ is defined in Proposition 19.

As a corollary of this statement, we have

$$\begin{aligned}
\bar{s}_i(\theta) &\stackrel{\text{def}}{=} \begin{bmatrix} p(1|y_i;\theta) \\ \ldots \\ p(g|y_i;\theta) \\ y_i \, p(1|y_i;\theta) \\ \ldots \\ y_i \, p(g|y_i;\theta) \end{bmatrix} = \mathsf{A}_{y_i} \begin{bmatrix} p(1|y_i;\theta) \\ \ldots \\ p(g|y_i;\theta) \end{bmatrix} \, , \\
\bar{s}(\theta) &\stackrel{\text{def}}{=} \begin{bmatrix} n^{-1} \sum_{i=1}^n p(1|y_i;\theta) \\ \ldots \\ n^{-1} \sum_{i=1}^n p(g|y_i;\theta) \\ n^{-1} \sum_{i=1}^n y_i \, p(1|y_i;\theta) \\ \ldots \\ n^{-1} \sum_{i=1}^n y_i \, p(g|y_i;\theta) \end{bmatrix} = \frac{1}{n} \sum_{i=1}^n \mathsf{A}_{y_i} \begin{bmatrix} p(1|y_i;\theta) \\ \ldots \\ p(g|y_i;\theta) \end{bmatrix} \, ,
\end{aligned} \tag{35}$$

where the probability $p(\cdot|y;\theta)$ is given by (34).

11.2 On the Assumption H3

Let A_y be the matrix defined in Proposition 19. It is proved in [12, Section 5] that $\mathsf{T}(s) \in \Theta$ if

$$s \in \mathcal{S} \stackrel{\text{def}}{=} \left\{ s = \frac{1}{n} \sum_{i=1}^{n} \mathsf{A}_{y_i} \rho_i, \rho_i = (\rho_{i,1}, \ldots, \rho_{i,g}) \in (\mathbb{R}_+)^g, \sum_{\ell=1}^{g} \rho_{i,\ell} = 1 \right\} .$$

The following statement shows that the SPIDER-EM sequence $\{\widehat{S}_k, k \geq 0\}$ is at least in

$$\widetilde{\mathcal{S}} \stackrel{\text{def}}{=} \left\{ s = \frac{1}{n} \sum_{i=1}^{n} \mathsf{A}_{y_i} \rho_i, \rho_i = (\rho_{i,1}, \ldots, \rho_{i,g}) \in \mathbb{R}^g, \sum_{\ell=1}^{g} \rho_{i,\ell} = 1 \right\} .$$

Proposition 22. *Assume that $\widehat{S}_{\text{init}} \in \mathcal{S}$. Then, for any $t \in \mathbb{N}$, $\mathsf{S}_{t,0} \in \mathcal{S}$ and for any $k \geq 0$, $\widehat{S}_{t,k} \in \widetilde{\mathcal{S}}$ and $\mathsf{S}_{t,k} \in \widetilde{\mathcal{S}}$.*

Proof. It is trivially seen from (35) that $\mathsf{S}_{t,0} \in \mathcal{S}$ for any $t \in \mathbb{N}$. Define $\rho_i^{(t,0)} \in (\mathbb{R}_+)^g$ and $\hat{\rho}_i^{(t,0)} \in (\mathbb{R}_+)^g$ such that

$$\mathsf{S}_{t,0} = \frac{1}{n} \sum_{i=1}^{n} \mathsf{A}_{y_i} \rho_i^{(t,0)} , \qquad \widehat{S}_{t,0} = \frac{1}{n} \sum_{i=1}^{n} \mathsf{A}_{y_i} \hat{\rho}_i^{(t,0)} ;$$

note that by (35), $\sum_{\ell=1}^{g} \rho_{i,\ell}^{(t,0)} = 1$ and by assumption, $\sum_{\ell=1}^{g} \hat{\rho}_{i,\ell}^{(t,0)} = 1$.

From line 5 of algorithm 1, we have when $k < k_{\text{in}} - 1$,

$$\mathsf{S}_{t,k+1} = \frac{1}{n} \sum_{i=1}^{n} \mathsf{A}_{y_i} \left(\rho_i^{(t,k)} + \frac{n}{\mathsf{b}} 1_{i \in \mathcal{B}_{t,k+1}} \left\{ p(\cdot|y_i; \mathsf{T}(\widehat{S}_{t,k})) - p(\cdot|y_i; \mathsf{T}(\widehat{S}_{t,k-1})) \right\} \right)$$

where $p(\cdot|y;\theta)$ is defined by (34), thus implying that

$$\rho_i^{(t,k+1)} = \rho_i^{(t,k)} + \frac{n}{\mathsf{b}} 1_{i \in \mathcal{B}_{t,k+1}} \left\{ p(\cdot|y_i; \mathsf{T}(\widehat{S}_{t,k})) - p(\cdot|y_i; \mathsf{T}(\widehat{S}_{t,k-1})) \right\} .$$

Hence by a trivial induction, $\sum_{\ell=1}^{g} \rho_{i,\ell}^{(t,k+1)} = 1$ for any $i = 1, \ldots, n$. From **??** and line 9 of algorithm 1, we have for any $k \geq 0$,

$$\widehat{S}_{t,k+1} = \frac{1}{n} \sum_{i=1}^{n} \mathsf{A}_{y_i} \left((1 - \gamma_{t,k+1}) \hat{\rho}_i^{(t,k)} + \gamma_{t,k+1} \rho_i^{(t,k+1)} \right)$$

thus implying that

$$\hat{\rho}_i^{(t,k+1)} = (1 - \gamma_{t,k+1}) \hat{\rho}_i^{(t,k)} + \gamma_{t,k+1} \rho_i^{(t,k+1)} .$$

Here again, by a trivial induction, we have $\sum_{\ell=1}^{g} \hat{\rho}_{i,\ell}^{(t,k+1)} = 1$ for any $i = 1, \ldots, n$. □

11.3 Numerical Analysis

11.3.1 The data set

We consider $n = 6 \times 10^4$ observations in $\mathbb{R}^p$, $p = 20$; modeled as independent observations from a mixture of Gaussian distributions with $g = 12$ components. These data are obtained from the MNIST data training set available at http://yann.lecun.com/exdb/mnist.

The set contains $n = 6 \times 10^4$ examples of size 28×28; among these pixels, 67 are constant over all the images and are removed yielding to observations of length 717. A PCA is performed in order to reduce the dimensionality to $p = 20$ features.

11.3.2 The algorithms

We compare EM, iEM, Online EM, FIEM and sEM-vr implemented as described in algorithm 2 to algorithm 6. The map T is given by Proposition 20.

The design parameters $\mathsf{b}, \gamma_{t,k+1}$ are fixed to

- $\mathsf{b} = 100$,
- for all the algorithms except iEM, the step size is constant and equal to $5\,10^{-3}$. In iEM, $\gamma_{k+1} = 1$.

Initialization. For all the algorithms and all the paths, the same initial value $\widehat{S}_{\text{init}}$ is considered. It is obtained as follows: we run the random initialization technique described in [19] in order to obtain $\theta_{\text{init}} \in \Theta$, and then we set $\widehat{S}_{\text{init}} \overset{\text{def}}{=} \bar{s}(\theta_{\text{init}})$. Below, $\widehat{S}_{\text{init}}$ is such that $-\,\mathrm{W}(\widehat{S}_{\text{init}}) = -58.3097$ (the constant term $p\log(2\pi)/2$ is omitted in this evaluation, and in any evaluation of the log-likelihood given below).

Mini-batch. The mini-batches are independent, and sampled at random in $\{1, \ldots, n\}$ with replacement. For a fair comparison of the algorithms, they share the same seed; another seed is used for FIEM which requires a second sequence of minibatches $\{\overline{\mathcal{B}}_{k+1}, k \geq 0\}$.

An epoch. In the analyses below, *an epoch* is defined as the selection of n examples:

- For EM, an epoch is one iteration $\widehat{S}_k \to \widehat{S}_{k+1}$. It necessitates the computation of n conditional expectations $\bar{s}_i$ and of a single optimization $\mathsf{T}(\widehat{S})$.
- For iEM and Online EM, an epoch is n/b iterations $\widehat{S}_k \to \widehat{S}_{k+1}$. It necessitates the computation of n conditional expectations $\bar{s}_i$ and of n/b optimizations $\mathsf{T}(\widehat{S})$.
- For FIEM, an epoch is n/b iterations $\widehat{S}_k \to \widehat{S}_{k+1}$. It necessitates the computation of $2n$ conditional expectations $\bar{s}_i$ and of n/b optimizations $\mathsf{T}(\widehat{S})$.
- For sEM-vr and SPIDER-EM, an epoch is either one iteration $\widehat{S}_{t,-1} \to \widehat{S}_{t,0}$ or n/b iterations $\widehat{S}_{t,k} \to \widehat{S}_{t,k+1}$ for $k < k_{\text{in}} - 1$. They resp. necessitate the computation of n and $2n/\mathsf{b}$ conditional expectations $\bar{s}_i$ and of 1 and n/b optimizations $\mathsf{T}(\widehat{S})$.

Hybrid methods. Since FIEM, sEM-vr and SPIDER-EM are variance reduction methods w.r.t. Online EM, we advocate to combine them with few steps of Online EM. Here, we start with $\text{kswitch} = 2$ epochs of Online EM and obtain $\widehat{S}_1, \widehat{S}_2$; before switching to FIEM, sEM-vr and SPIDER-EM.

Value of $k_{\max}$. The number $k_{\max}$ is fixed in order to compare the algorithms with the same number of epochs equal to 150. For EM, $k_{\max} = 150$; for Online EM and iEM, $k_{\max} = 150\, n/\mathsf{b}$; for FIEM, $k_{\max} = (150 - \text{kswitch})\, n/\mathsf{b}$; for sEM-vr, $k_{\text{out}} = (150 - \text{kswitch})/2$ and $k_{\text{in}} = 1 + n/\mathsf{b}$; and for SPIDER-EM, $k_{\text{out}} = (150 - \text{kswitch})/2$ and $k_{\text{in}} = 1 + n/\mathsf{b}$.

11.3.3 Experimental Results

We first analyze the behavior of the functional W along a path of the algorithm. We display on Figure 4 a Monte Carlo approximation, computed from 40 independent runs, of the expectation of the normalized log-likelihood as a function of the number of epochs. Different algorithms are considered: EM remains trapped in a local extremum while the stochastic EM algorithms succeed in exiting to a better limiting point. Online EM is far more variable than iEM, FIEM, sEM-vr and SPIDER-EM. The convergence of iEM is longer, when compared to FIEM, sEM-vr and SPIDER-EM.

On Figure 5 and Figure 6, for each of the algorithms FIEM, sEM-vr and SPIDER-EM, four different realizations of a path of the normalized likelihood are displayed as a function of the number of epochs. These four sets of curves differ from the selection of the sequence of mini-batches. The staircase behavior of the paths of sEM-vr and SPIDER-EM comes from the two successive kinds of epoch: one corresponds to a single optimization and a full scan of the data set and the other one corresponds to n/b optimizations and the use of n/b minibatches; the largest increase of W corresponds to the second type of epoch. Based on this criterion, the three algorithms are equivalent.

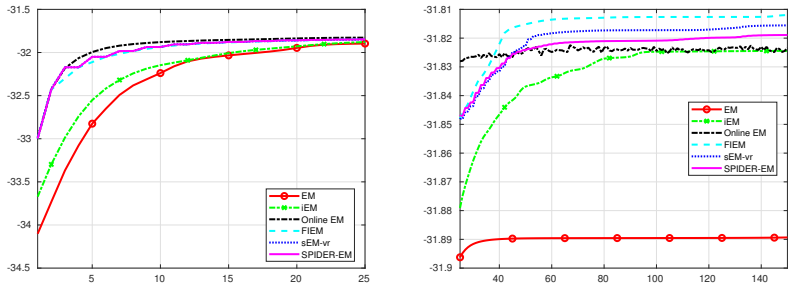

Figure 4: Monte Carlo approximation (computed over 40 independent runs) of $-\mathbb{E}[\mathrm{W}(\widehat{S}_\ell)] = -\mathbb{E}[F \circ \mathsf{T}(\widehat{S}_\ell)]$ against the number of epochs. [left] Epochs 1 to 25; [right] epochs 25 to 150.

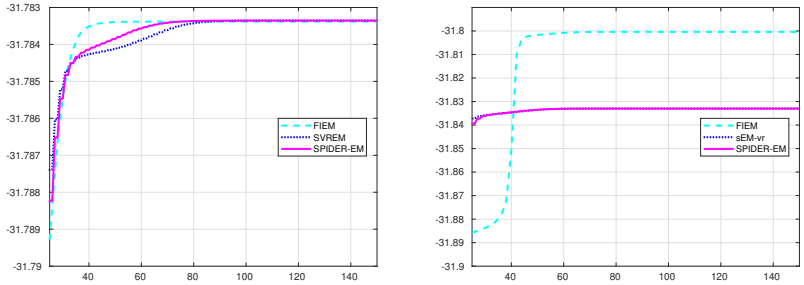

Figure 5: The objective function $-\mathrm{W}(\widehat{S}_\ell) = -F \circ \mathsf{T}(\widehat{S}_\ell)$ against the number of epochs along two (left, right) independent runs of `FIEM`, `sEM-vr` and `SPIDER-EM`. The first 25 epochs are discarded.

Figure 7 displays the evolution of the $g = 12$ iterates $\{\alpha_1, \ldots, \alpha_g\}$ along a path of many algorithms. Figure 8 display the evolution of the $p = 20$ eigenvalues of the covariance matrix Σ along a path of many algorithms. Here again, we observe a strong variability of `Online EM` when compared to the other algorithms.

Figure 9 and Figure 10 display 40 independent realizations of the squared norm of the mean field h as a function of the number of epochs for different algorithms. It may be seen that `Online EM` has a strong variability and `FIEM`, `sEM-vr`, `SPIDER-EM` succeed in reducing this variability. `FIEM` converges more rapidly than `iEM`, and they achieve the same level of accuracy (here not better than 10^{-6}). `sEM-vr` and `SPIDER-EM` have the same level of accuracy, which is most often far smaller than the one reached by `FIEM` (more than 75% of the paths reached an accuracy level of 10^{-10} after 150 epochs). Based on this criterion, we will definitively advocate the use of `sEM-vr` or `SPIDER-EM` when compared to `iEM`, `Online EM` and `FIEM`.

Figure 11 and Figure 12 display the boxplots of 40 independent realizations of $\|h(\widehat{S}_\ell)\|^2$ at time in $\{20, 40, 60, 80, 110\}$ epochs for different algorithms. In Figure 12, `Online EM` is not displayed since it is too large (compare the third plot on Figure 11 and the first one on Figure 12). The quantities $\{\|h(\widehat{S}_\ell)\|^2, \ell \geq 0\}$ are the key informations for deriving the complexity bounds in Theorem 2. The plots below show again that for small, medium and large values of the number of epochs k, `sEM-vr` and `SPIDER-EM` provide the best results.

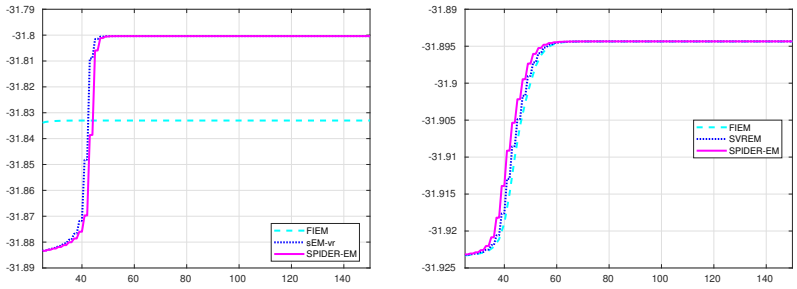

Figure 6: The objective function $-\mathrm{W}(\widehat{S}_\ell) = -F \circ \mathsf{T}(\widehat{S}_\ell)$ against the number of epochs along two (left,right) independent runs of `FIEM`, `sEM-vr` and `SPIDER-EM`. The first 25 epochs are discarded.

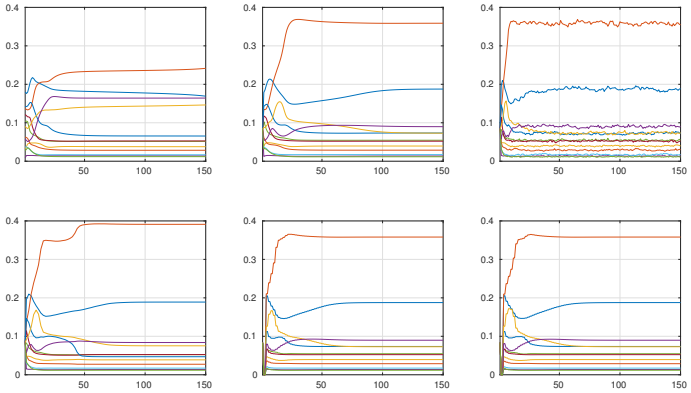

Figure 7: Evolution of the $g = 12$ iterates $\alpha_k = (\alpha_{k,1}, \ldots, \alpha_{k,g})$ against the number of epochs, for `EM`, `iEM` and `Online EM` on the top from left to right; `FIEM`, `sEM-vr` and `SPIDER-EM` on the bottom from left ro right.

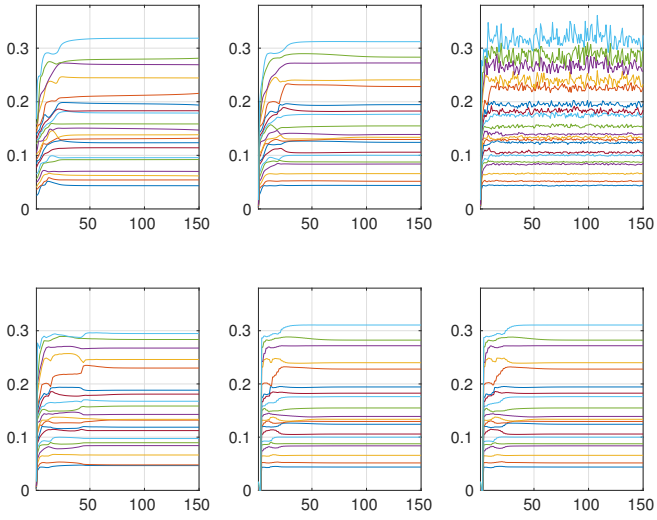

Figure 8: Evolution of the $p = 20$ eigenvalues of the iterates $\{\Sigma_\ell, \ell \geq 0\}$ against the number of epochs ℓ, for `EM`, `iEM` and `Online EM` on the top from left to right; `FIEM`, `sEM-vr` and `SPIDER-EM` on the bottom from left ro right.

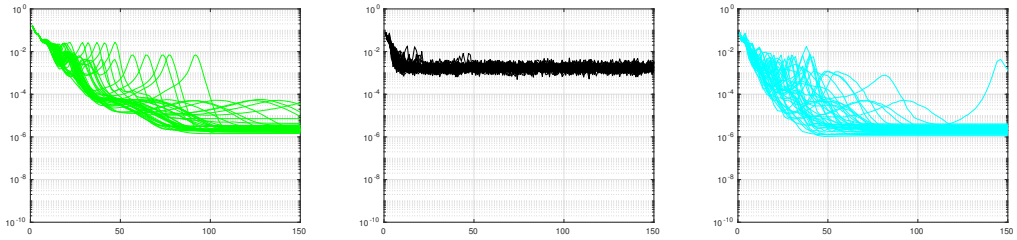

Figure 9: [left] We display 40 independent realizations of the squared norm of the mean field $\ell \mapsto \|h(\widehat{S}_\ell)\|^2$ as a function of the number of epochs, along a `iEM` path. [center] same analysis for `Online EM`. [right] same analysis for `FIEM`.

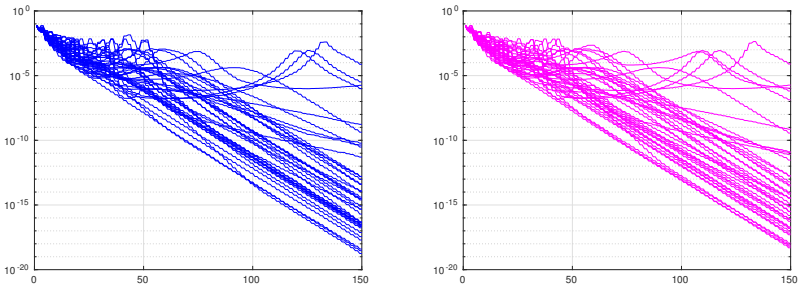

Figure 10: [left] We display 40 independent realizations of the squared norm of the mean field $\ell \mapsto \|h(\widehat{S}_\ell)\|^2$ as a function of the number of epochs, along a `sEM-vr` path. [right] same analysis for `SPIDER-EM`.

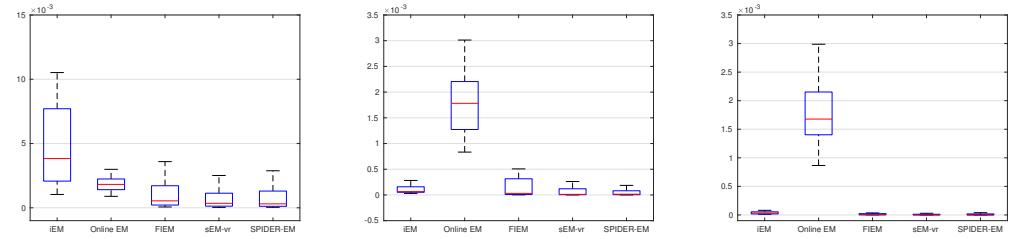

Figure 11: Boxplots of 40 independent points of $\|h(\widehat{S}_\ell)\|^2$ [left] at time 20 epochs; [center] at time 40 epochs; [right] at time 60 epochs. The outliers are removed.

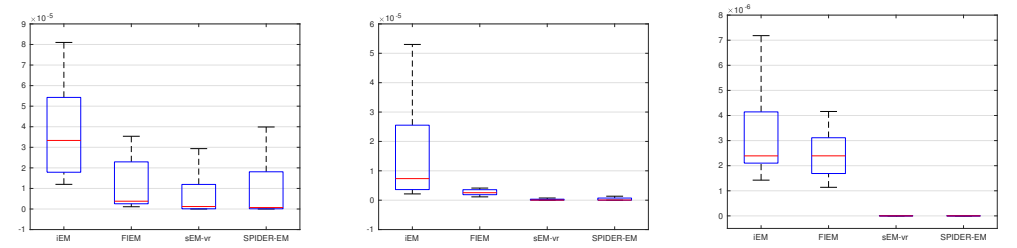

Figure 12: Boxplots of 40 independent points of $\|h(\widehat{S}_\ell)\|^2$ [left] at time 60 epochs; [center] at time 80 epochs; [right] at time 110 epochs. The outliers are removed.